WIRELESS COMPUTING IN MEDICINE

Nature Inspired Computing Series

Editors:
Mary Mehrnoosh Eshaghian-Wilner
and Albert Zomaya

Eshaghian-Wilner—Bio-Inspired and Nanoscale Integrated Computing
Dorronsoro—Evolutionary Algorithms for Mobile Ad hoc Networks
Bouvry—Evolutionary Algorithms for Mobile Ad hoc Networks

WIRELESS COMPUTING IN MEDICINE

From Nano to Cloud with Ethical and Legal Implications

Edited by

MARY MEHRNOOSH ESHAGHIAN-WILNER

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication data applied for

ISBN: 9781118993590

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

CONTRIBUTORS

Adnane Abdelghani, Nanotechnology Laboratory, National Institute of Applied Science and Technology, Carthage University, Charguia Cedex, Tunisia

Nikita Ahuja, EE-Systems, University of Southern California, Los Angeles, CA, USA

Murali Annavaram, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

Katie Atalla, EE-Systems, University of Southern California, Los Angeles, CA, USA

Ido Bachelet, Faculty of Life Sciences and The Institute of Nanotechnology and Advanced Materials, Bar Ilan University, Ramat Gan, Israel

William Sims Bainbridge, Division of Information and Intelligent Systems, National Science Foundation, Arlington, VA, USA

Saeid Barzegarjalali, EE-Systems, University of Southern California, Los Angeles, CA, USA

Jessica Beltran-Marquez, Instituto Tecnológico de Sonora, ITESM, Universidad La Salle, Mexico City, DF, Mexico

Gaymon Bennett, School of Politics and Global Studies, Arizona State University, Tempe, AZ, USA

Amber Bhargava, EE-Systems, University of Southern California, Los Angeles, CA, USA

Luis A. Castro, Spectrolab, Inc., Sylmar, CA, USA

Ayush Chaudhary, Intel, Santa Clara, CA, USA

Edgar Chavez, Department of Computer Sciences, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Zhaoqi Chen, EE-Systems, University of Southern California, Los Angeles, CA, USA

Janet Cheung, EE-Systems, University of Southern California, Los Angeles, CA, USA

Franceli L. Cibrian, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

T. Soren Craig, EE-Systems, University of Southern California, Los Angeles, CA, USA

Lizbeth Escobedo, National Institute of Technology of Mexico, Mexico City, Mexico

Mary Mehrnoosh Eshaghian-Wilner, EE-Systems, University of Southern California, Los Angeles, CA, USA

Jesus Favela, Department of Computer Sciences, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Zhuochen Ge, EE-Systems, University of Southern California, Los Angeles, CA, USA

Nikila Goli, EE-Systems, University of Southern California, Los Angeles, CA, USA

Kalyani Gonde, EE-Systems, University of Southern California, Los Angeles, CA, USA

Rakesh Guduru, College of Engineering and Computing, Florida International University, Miami, FL, USA

Arushi Gupta, EE-Systems, University of Southern California, Los Angeles, CA, USA

J. Benjamin Hurlbut, School of Politics and Global Studies, Arizona State University, Tempe, AZ, USA

Sakhrat Khizroev, College of Engineering and Computing, Florida International University, Miami, FL, USA

Rebecca Lee, EE-Systems, University of Southern California, Los Angeles, CA, USA

Sangwon Lee, Computer Science Department, University of Southern California, Los Angeles, CA, USA

Wan Lee, Mechanical and Aerospace Engineering, California State University, Long Beach, CA, USA

Renjun Liu, EE-Systems, University of Southern California, Los Angeles, CA, USA

Shuping Liu, eBay, Inc., San Jose, CA, USA

José J. López, Sociological and Anthropological Studies, University of Ottawa, Ottawa, ON, Canada

Noam Mamet, Faculty of Life Sciences and The Institute of Nanotechnology and Advanced Materials, Bar Ilan University, Ramat Gan, Israel

Sylvain Martel, Department of Computer Engineering, Polytechnique Montréal, Montreal, QC, Canada

Raj Mehta, EE-Systems, University of Southern California, Los Angeles, CA, USA

Bouzid Menaa, Department of Infection Control, Hymetec SA, Isnes, Belgium

Farid Menaa, Department of Nanomedicine, Oncology and Stem Cells, Fluorotronics, Inc., San Diego, CA, USA

José Mercado, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Clark A. Miller, School of Politics and Global Studies, Arizona State University, Tempe, AZ, USA

Urbashi Mitra, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

Fredrick Japhet Mtenzi, School of Computing, College of Sciences and Health, Dublin Institute of Technology, Dublin, Ireland

Shiva Navab, Broadcom, Irvine, CA, USA

Rene Navarro, Department of Industrial Engineering, University of Sonora, Hermosillo, SO, Mexico

Adarsh Nayak, EE-Systems, University of Southern California, Los Angeles, CA, USA

Mathieu Noury, Sociological and Anthropological Studies, University of Ottawa, Ottawa, ON, Canada

Anand Panangadan, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

Alice C. Parker, EE-Systems, University of Southern California, Los Angeles, CA, USA

Alekhya Sai Nuduru Pati, EE-Systems, University of Southern California, Los Angeles, CA, USA

Sukanya Patil, IIT Bombay, Mumbai, Maharashtra, India

Moises Perez, Department of Computer Sciences, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Linda Powers, Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA

Andrew Prajogi, EE-Systems, University of Southern California, Los Angeles, CA, USA

Eduardo Quintana, Department of Computer Sciences, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Kodiak Ravicz, EE-Systems, University of Southern California, Los Angeles, CA, USA

Cauligi S. Raghavendra, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

Kui Ren, Department of Computer Science and Engineering, University of Buffalo, State University of New York, Buffalo, NY, USA

Marcela Rodriguez, Department of Computer Engineering, Autonomous University of Baja California, Mexicali, BC, Mexico

Heather M. Ross, School of Politics and Global Studies, Arizona State University, Tempe, AZ, USA

Rakshith Saligram, EE-Systems, University of Southern California, Los Angeles, CA, USA

Gaurav Sarkar, EE-Systems, University of Southern California, Los Angeles, CA, USA

Mike Schlesinger, EE-Systems, University of Southern California, Los Angeles, CA, USA

Yesha Shah, EE-Systems, University of Southern California, Los Angeles, CA, USA

Umang Sharma, EE-Systems, University of Southern California, Los Angeles, CA, USA

Ben Shiroma, EE-Systems, University of Southern California, Los Angeles, CA, USA

Ashit Talukder, Information Access Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

Monica Tentori, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico

Pujal Trivedi, EE-Systems, University of Southern California, Los Angeles, CA, USA

Abhishek Uppal, EE-Systems, University of Southern California, Los Angeles, CA, USA

Sandeep Kumar Vashist, Hahn-Schickard-Gesellschaft für Angewandte Forschung E.V. (HSG-IMIT), Freiburg, Germany

Janet Meiling Wang-Roveda, Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA; University of Southern California, Los Angeles, CA, USA

Xiaojun Xian, Center for Bioelectronics and Biosensors, The Biodesign Institute, Arizona State University, Phoenix, AZ, USA

Kai Xie, School of Aerospace Science and Technology, Xidian University, Xi'an, Shaanxi, China

Shu Han Wu, EE-Systems, University of Southern California, Los Angeles, CA, USA

Kun Yue, EE-Systems, University of Southern California, Los Angeles, CA, USA

Daphney-Stavroula Zois, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA

FOREWORD

Wireless Computing in Medicine: From Nano to Cloud with Its Ethical and Legal Implications (edited by Professor Mary Mehrnoosh Eshaghian-Wilner) is an exciting book that deals with a wide range of topical themes in the field of healthcare and biomedicine. The book also probes the legal and ethical issues that are of immense importance in health and medicine.

Today, we are witnessing many advances in healthcare brought about by the impact of computing on the practice of medicine and biomedical sciences. Many of these advances are due to the developments in algorithmics, wireless networking, high-performance computing, and many others. The book also showcases some of the ethical and legal issues that cannot be ignored when marrying computing technology with medicine and health sciences.

I believe that the current book is a great addition to the literature. It will serve as a source of up-to-date research in this continuously evolving area. The book also provides an opportunity for researchers to explore the use of advanced computing technologies and their impact on enhancing our capabilities to conduct more sophisticated studies.

The book should be well received by the research and development community and can be beneficial for graduate classes focusing on biomedical engineering, bio-nanotechnology, and health informatics.

Finally, I would like to congratulate Dr. Eshaghian-Wilner for a job well done, and I look forward to seeing the book in print.

Albert Y. Zomaya
Sydney, January 2016

PREFACE

I recently celebrated my 50th birthday 26 productive years after I received my Ph.D. On this important milestone, I reflected back on my life, as I could not help but find myself in total agreement with what both Aristotle and Einstein said: the more one learns or knows, the more one realizes how much he/she does not know. I always wanted to learn more, so over the years I have expanded my parallel processing expertise from heterogeneous computing (topic of my first book) to bio-inspired and nanoscale integrated computing (topic of my second book). Expanding the application of both of these technologies further to medicine with an emphasis on their legal and ethical aspects is the main aim of this third book. This book is a product of the progression of my research, from undergraduate study until now.

In the early part of my research career and as a research student at the University of Southern California (USC), I concentrated on the design of efficient very-large-scale integration (VLSI) architectures and parallel algorithms, especially for image and signal processing. Such research focused on my development of fast algorithms for solving geometric problems on the Mesh-of-Trees architecture. These techniques have been applied to several other architectures, including bus-based architectures and later on architectures such as the systolic reconfigurable mesh. Thirty years since their inception, these results are still showing their utility in the design of graphics processing unit (GPU) architectures.

Later, as part of my Ph.D., I focused my attention on applying my Mesh-of-Trees results to the area of optical computing. I produced the Optical Model of Computation (OMC) model, through which I was able to show the computational limits and the space–time tradeoffs for replacing electrical wires with free-space optical beams in VLSI chips. Based on the model, I designed several generic electrooptical architectures, including the electrooptical crossbar design that includes a switching speed in

the order of nanoseconds. This design was later extended to an architecture called "optical reconfigurable mesh" (ORM). Algorithms designed on ORM have a very fast running time because ORM comprises a reconfigurable mesh in addition to having both a microelectromechanical system (MEMS) and electrooptical interconnectivity. OMC is a well-referenced model that has been shown to have superior performance compared to many other parallel and/or optical models. Based on OMC, the well-known local memory parallel random access memory (PRAM) model was developed. Furthermore, variations of OMC were adopted by the industry in designing MEMS chips.

Soon after I graduated, I took a leading role in starting the heterogeneous computing field. I am the editor of the field's first book, *Heterogeneous Computing*, and the cofounder of the IEEE Heterogeneous Computing Workshop. The book in conjunction with the workshop shaped the field and paved the path to today's "cloud computing." As one of the first paradigms for executing heterogeneous tasks on heterogeneous systems, I developed the Cluster-M model. Prior models such as PRAM and LogP each had their limitations because they could not handle arbitrary systems or structures with heterogeneous computing nodes and interconnectivity. Cluster-M mapping is still the fastest known algorithm for mapping arbitrary task graphs onto arbitrary system graphs.

For over a decade now, I have been focusing on the bio- and nanoapplications of my work. I am a founding series coeditor of "Nature-Inspired Computing" for John Wiley & Sons and have edited the first book of this series, *Bio-inspired and Nanoscale Integrated Computing*. This is truly a multidisciplinary topic that required a significant amount of training from several fields. Toward this multidisciplinary field, I have studied various techniques for designing nanoscale computing architectures where computations are subject to quantum effects. One of the most notable works I have produced in this area is a joint work with my colleagues at the University of California, Los Angeles (UCLA). The work was announced as a breakthrough result by numerous media outlets and was explained in many review articles worldwide. It involved the design of a set of highly interconnected multiprocessor chips with spin waves. These designs possess an unprecedented degree of interconnectivity that was not possible previously with electrical VLSI interconnects, because they can use frequency modulation to intercommunicate among nodes via atomic waves. Furthermore, the information is encoded into the phase of spin waves and is transferred through ferromagnetic buses without any charge transmission. The spins rotate as propagating waves, and as such, there is no particle (electron/hole) transport. This feature results in significantly lower power consumption as compared to other nanoscale architectures.

Extending nanoscale computing to cellular biology, I have studied applications of spin-wave architectures for DNA sequence matching. I have shown that these designs have a superior algorithmic performance for such applications. Also, because they can operate at room temperature, they have a great potential to be used as part of miniature implantable devices for biomedical and bio-imaging applications. I have been investigating efficient methods for designing injectable nanorobots that can be used for the detection and treatment of various diseases, especially cancer.

My most recent area of research has been in technology law. I am investigating how various forms of emerging technologies may be impacted by, and come into conflict with US and international policies and laws. For example, while pervasive (heterogeneous/ubiquitous) computing and nanotechnology are two technologies that are entirely different from each other, they both are seemingly invisible: one in terms of interconnectivity and the other in terms of size. Their overwhelming potential coupled with their peculiar nature can continuously magnify challenges to policies and laws that protect rights and property. Such challenges and related legal and ethical issues are discussed further as a chapter in this book, especially as applied to their applications in wireless computing for medicine.

This book contains 21 chapters presented in five parts. In Part 1, my students and I have presented an introduction to the book in the first chapter, and in the second chapter, we have given an introduction to the two wireless technologies used in the book: pervasive computing and nanocomputing.

In Part 2—Pervasive Wireless Computing in Medicine, there are seven chapters detailing pervasive computing for medicine. Authored by the leading scientist in the field, these chapters cover a wide range of topics such as pervasive computing in hospitals, diagnostic improvements: treatment and care, collaborative opportunistic sensing of human behavior with mobile phone, pervasive computing to support individuals with cognitive disabilities, wireless power for implantable devices, energy-efficient physical activity detection in wireless body area networks, and Markov decision process for adaptive control of distributed body sensor networks.

Similarly, in Part 3—Nanoscale Wireless Computing in Medicine, there are seven chapters authored by leading scientists. These chapters all focus on the application of nanocomputing in medicine. The topics include an introduction to nanomedicine, nanomedicine using magneto-electric nanoparticles, DNA computation in medicine, graphene-based nanosystem for the detection of proteomic biomarkers of disease: implication in translational medicine, modeling brain disorders in silicon nanotechnologies, linking medical nanorobots to pervasive computing, and nanomedicine's transversality: some implications of the nanomedical paradigm.

Finally, in Part 4—Ethical and Legal Aspects of Wireless Computing in Medicine, the ethical and legal aspects of wireless computing in medicine are presented in four chapters by leading scholars in this area. The topics include ethical challenges of ubiquitous health care, ethics of ubiquitous computing in health care, privacy protection of electronic healthcare records in e-healthcare systems, and ethical, privacy, and intellectual property issues in nanomedicine. After this third section, we provide a brief conclusion in Part 5.

In addition to the two introductory chapters and the conclusion chapter, I have coauthored one of the chapters in Part 1 ("Wireless Power for Implantable Devices"), two of the chapters in Part 2 ("An Introduction to Nanomedicine and Nanomedicine Using Magneto-electric Nanoparticles"), and one chapter in part 3 ("Ethical, Privacy, and Intellectual Property Issues in Nanomedicine"). Evidently, most of these chapters deal specifically with nanomedicine. I believe nanomedicine is one of this century's most promising scientific fields in which we can soon expect to see many

life-altering advancements. Targeted delivery of drugs to cancer cells is already in animal-testing stages with very impressive preliminary results. Furthermore, various techniques are currently being studied to develop nanorobots that can aid in both detection and treatment of cells.

I invite you to learn more about this exciting field by reading this book. You will see that the more you learn, the more you will realize how much there is yet to be learned and discovered in this growing and fascinating field.

And finally, I would like to take this opportunity to thank all of those who made this book possible. First and foremost, I would like to express my gratitude to Professor Albert Zomaya who has been coediting Wiley's Nature-Inspired Computing series with me. Next, my most sincere appreciation extends to all my distinguished co-authors who have so greatly contributed to this volume, and to the numerous other wonderful people at Wiley and USC, especially my students who have worked with me on this book.

I dedicate this book to my family. My parents, Mehdi and Molly, have always encouraged and helped me advance my education and career, and I am forever indebted to them. I am also beholden to my husband, Arthur, for being my best friend and advisor. My interest in the field of technology law was sparked by his legal expertise and our related discussions. Additionally, I am thankful to my brothers, Michael and Mark; my sister, Maggie, and her husband, Kamran; and my nephews, Jonathan and Justin, for providing so much joy and inspiration. But of the utmost importance, I am singularly appreciative of my remarkable daughter, Ariana Shaina, who has so brilliantly illuminated my life with her precious love, astonishing wisdom, and fiercely strong values. I couldn't be prouder of her.

Thank you for reading this book. Enjoy!

Mary Mehrnoosh Eshaghian-Wilner

PART I

INTRODUCTION

1

INTRODUCTION TO WIRELESS COMPUTING IN MEDICINE

Amber Bhargava, Mary Mehrnoosh Eshaghian-Wilner, Arushi Gupta, Alekhya Sai Nuduru Pati, Kodiak Ravicz, and Pujal Trivedi*

EE-Systems, University of Southern California, Los Angeles, CA, USA

1.1 INTRODUCTION

Constant population growth has increased the need for more advanced scientific solutions for ever-growing healthcare demands. It requires a new paradigm and technology for more effective solutions. There has been a booming growth in technology, which has resulted in devices becoming progressively smaller and more powerful. One result of computer technology advancing at exponential speeds is wireless computing, which combines current network technologies with wireless computing, voice recognition, Internet capability, and artificial intelligence, to create an environment where the connectivity of devices is unobtrusive and always available. But as this connectivity improves, so does the collection and retrieval of data. In the field of medicine, because hospitals collect large amounts of unnecessary data on patients, it is difficult for doctors to distinguish a real emergency. We need to improve the standard of medical care provided to patients by helping doctors make more informed decisions. Doctors also require a greater degree of accuracy while treating chronic diseases, or for example, treating cancerous cells without affecting the regular ones. This chapter aims at promoting the discussion on how the use of wireless computing in nanomedicine helps integrate health monitoring and healthcare more seamlessly in

*Authors are listed alphabetically by last name, as opposed to by order of contribution.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

the healthcare sector, ways it can help us to tackle the critical challenges faced by doctors and patients regardless of space and time, and also present cutting-edge perspectives and visions to highlight future developments.

Nanomedicine has been considered a possibility ever since the concept of nanotechnology was first articulated in 1959 by Richard Feynman, in his famous Caltech talk, "There's Plenty of Room at the Bottom." Feynman mentions that a friend of his says "You put the mechanical surgeon inside the blood vessel and it goes into the heart and 'looks' around." The application of nanomedicine has a strong potential for shifting myriad paradigms in the field of medicine. This is because nanomedicine operates at the molecular, organellar, and cellular levels; precisely where disease processes find their genesis. Once matured, these capacities will have immense benefits in terms of positive patient outcomes and the alleviation of human suffering across the board. There is a rapidly growing global trend toward the development of more compact, minimally invasive, intelligent, more accurate, and efficacious medical technologies. But the general consensus is that despite encouraging signals and growth, the field of nanomedicine has yet to fully mature.

Wireless computing is one of the techniques to help nanomedicine grow at the rate seen by the visionaries. It's a sure sign that wireless computing has entered a new era—in some ways even more telling than the PC's dominance or wireless communication's emergence. Wireless health system encompasses new types of sensing and communication of health information as well as new types of interactions among health providers and people, among patients and researchers, and among patients and corporations.

The convergence of two domains of current research—nanotechnology and distributed computing—presents a lot of applications in the field of medicine. In this chapter, we briefly summarize and present the technologies underlying the state-of-the-art research in the interdisciplinary field of medical wireless computing. In the first section, distributed computing, we discuss its usage in treating cognitive disabilities like Alzheimer's, autism, etc., and how it can increase the portability for monitoring the patient and reduce the redundancy of data. We also talk about the role of wireless power and Markov decision process (MDP) in distributed computing. In the second section, nanomedicine, we discuss about the technologies of nanocomputing and the ways they can be utilized in medicine. We also explain how we can model brain disorders and detect biomarkers using nanotechnology. In the third section, we discuss about ethics, privacy, and legal issues in the domain of nanomedicine, and how we can implement these in a safe, ethical way to gain benefits.

This chapter intends to provide readers with a sense of the breadth and depth of the field of wireless computing, and its potential effects on medicine. We define the technologies of wireless computing, from the software that run cloud computing data centers, to the technologies that allow new sensors to work. We also provide readers with case studies of how these technologies are being implemented in the medical field through both integrating into current systems and creating new forms of medical applications.

We hope that this chapter will be useful to anyone who wishes to learn about the interdisciplinary field of wireless computing. We have tried our best to make this material understandable at a beginner level. Students with backgrounds in the fields

of medicine, computing, health informatics, and even public policy should be able to understand the material presented within and gain useful insights.

1.2 DEFINITION OF TERMS

Nanomedicine: It is the application of nanotechnology (the engineering of tiny machines) to the prevention and treatment of disease in the human body. This evolving discipline has the potential to dramatically change medical science. Nanomedicine is also defined as the monitoring, repair, construction, and control of human biological systems at the molecular level, using engineered nanodevices and nanostructures.

Healthcare workers: In this case, we use the term "healthcare workers" to describe the ecosystem of doctors, nurses, hospital administrators, and people in the government who are involved in healthcare policy.

Distributed computing: The technology that enables computing in devices and allows their communication wirelessly. This includes the infrastructure of distributed computing, including but not limited to the physical infrastructure of the Internet, the algorithms for directing traffic on the Internet, cell phones, and other "smart" devices.

Pervasive computing: From a technology standpoint, pervasive computing is mostly the same thing as distributed computing, but it also refers to the way that computing has become part of the fabric of our social existence. Pervasive computing refers to the idea that computing is invisible and everywhere, that it is a part of our daily lives that we take for granted.

1.3 BRIEF HISTORY OF WIRELESS HEALTHCARE

Healthcare is a remarkably interdisciplinary field. Biology, chemistry, immunology, and psychology are just a few of the skills that are necessary for healthcare officials to understand in order to produce a working healthcare system.

An important part of any science is the gathering of information; this is no less true in healthcare. John Snow, the father of modern epidemiology, discovered the source of the nineteenth-century cholera outbreak in London by creating a map of all known cholera cases (www.udel.edu/johnmack/frec682/cholera/), finding patterns in the congregated data to help treat cholera not as the disease affecting a particular patient, but rather treating cholera as a pervasive condition that affected the City of London.

John Snow and other doctor-scientists ushered in a completely new form of medicine—the modern medicine that we have today. Epidemiology, a completely new field of medicine, led to discoveries such as the link between tobacco smoking and cancer, and along with the "germ theory" of medicine helped promote the use of disinfectants, which greatly improved our quality of life and increased the expected lifespan (www.sciencemuseum.org.uk/broughttolife/techniques/germtheory.aspx).

By looking at the total picture, congregating the data from many patients across London, John Snow was able to discover something about the nature of cholera that was completely invisible to anyone who only looked at each patient individually.

The information existed before, but defining epidemiology as a part of healthcare, as a new way to look at healthcare, gave doctors a revolutionary way to treat a disease.

Today, information has never been easier to gather or transmit. The Internet allows data to be sent almost instantaneously across the globe, and an enormous array of sensors are able to take our temperature, track our activity levels and sleep patterns, and gather other statistics about our lives. There are many new methods to gather information about patients and diseases that could drive the next revolution in medicine. These new methods are sometimes referred to as "pervasive computing."

The intention of this chapter is to introduce readers to the application of wireless computing in medicine by giving an overview of wireless computing itself, providing a few examples of wireless computing in action through contemporary research projects, which are pushing the edge of medicine, and defining wireless computing as a hardware as well as a software phenomenon by providing examples of new nanoscale technologies. We also want to provide our readers with an idea of how wireless computing is changing medicine through these improvements in information transmission and generation. How will our medical institutions, our doctor–patient relationships, our society as a whole, change because of wireless computing in medicine? Although wireless computing may bring great improvements to our medical system, it also poses certain threats, such as information security and the fear of a surveillance state. These questions and more are discussed.

1.4 WHAT IS WIRELESS COMPUTING?

Wireless computing has gone by many names including ubiquitous computing, distributed computing, heterogeneous computing, physical computing, and the Internet of Things. What separated wireless computing from traditional computing was the idea that any object in the world—not just a desktop computer or set of servers—could do computing.

In this chapter, we will view wireless computing in two parts. One is the distributed computing architecture that allows applications to connect across the Internet. This includes all parts of the system that run in the "back end," from servers to software. The other area that we will consider is the ever-growing list of "smart" things that collect data or user input and upload to the overlaying structure. These devices are not only increasing the amount of data that we can collect but are also changing the ways that we collect such data. Breakthroughs in nanotechnology may allow sensors to become truly invisible and computing to become truly wireless, around us at all times. We consider the ways in which new forms of both connectivity and data generation are changing the way wireless computing works, extending its reach.

Today, wireless computing has been integrated into many facets of our lives so successfully that we barely stop to wonder at the fact that our phones will tell us when it's time to pay the electric bill; we take for granted that our homes will notify the police when there has been a break-in, and we fully expect that our friends will be instantly aware of our new high score on the latest game. However, when we

consider applying wireless computing to medicine, there are a number of questions that remain unanswered.

Medical applications entail a unique set of constraints and needs, from an increased desire for information security to the need for very accurate data. However, we believe that the application of wireless computing to medicine could provide many benefits to our healthcare system (in this chapter we primarily discuss the healthcare system of the United States, but we also look at research that is being done in Europe and across the globe). We discuss wireless computing in detail in Chapter 2, this volume [1]. We discuss both applications in distributed computing and in end-user devices. Some of the technologies are currently implemented, and we use these as case studies for implementing these solutions in a broader context; others have yet to be implemented but have strong potential for revolutionizing the way that we treat disease.

1.5 DISTRIBUTED COMPUTING

There are many applications for distributed computing in medicine. Analyzing the "big data" of medicine could allow doctors to prescribe truly personalized medicine, while the application of various sensors allows for a greater degree of accuracy in treating chronic disease. Mobile and wireless technology can improve the standard of medical care being provided to patients and will also help them make more informed decisions regarding their health.

In the area of distributed systems, much of the technology is already established. The architecture of the Internet, although nebulous, is fairly well defined. The architecture for dealing with large sets of data exists and is being used in other nonmedical applications; instead, it is the particular implementation of these technologies to medical uses that is interesting. Medical applications have various restrictions that are different from other purposes—the privacy of information is especially important and is one of the more talked-about issues, but there also are issues of patient safety—all devices used by medical professionals should be held to a higher standard than devices used by the average consumer. These issues pose technical challenges for both hardware devices and software algorithms.

Take for example an alert system for patients in a hospital or a nursing home that would tell nurses when a patient needs assistance, or emergency care. Theoretically, this system could help nurses be more efficient and respond to emergencies with a shorter delay. From a technical perspective, the challenges include but are not restricted to the following:

1. Create a set of sensors that can detect an anomaly in the patient that requires nurse assistance.
2. Network these sensors to a database.
3. Provide a user interface to the nurses or other health professionals who will be using the device.

4. Assure that all alerts will be delivered with some maximum time delay.
5. Make sure that the patient's data is secure.

None of these challenges are trivial and most require knowledge of both the technology and the medical application. For example, the Internet Protocol (TCP/IP) does not guarantee a minimum delay for the delivery of information packets. If an application cannot tolerate small delays, TCP/IP might not be the best method with which to implement the system. Instead, an Intranet, or self-contained Internet that does not interface with the larger Internet, may be a better option, and in that system you may be able to ensure that information is delivered within a certain amount of time. On the other hand, user interfaces may seem to be a fairly simple thing, which many individuals and companies create on a daily basis, but when it comes to medical applications, it is important to create an interface that is not only easy to use but also prioritizes alerts that are important and at the same time does not swamp health professionals with a stream of data that is bothersome and takes away from their other activities. An overview of the distributed computing from a technical perspective is included in Chapter 2, this volume [1].

Each application of distributed computing faces its own unique set of challenges. Applying distributed computing to hospitals is very different from the quantified self-movement where individuals track their own blood pressure or activity at home, and the quantified self-movement is different from applications where medical professionals are trying to use home-collected data in a professional way.

One effect of distributed systems in medicine is to extend the reach of medicine from the doctor's office or hospital into the home, and to extend the definition of medicine from reactive cures to preventative actions. Distributed computing not only allows the doctors to reach their patients in the home, but is also fundamentally changing the way that doctors and patients interact. Technology has continuously been changing how people interact with everything; in medicine, you see this in simple things that we take for granted such as looking up a disease on WebMD, being able to schedule a doctor's appointment over the phone, or call 911 for an emergency. Even complicated technical procedures such as remote surgeries are possible today (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422462/). Wireless computing continues to develop new ways for doctors and patients to interact, through ever-more-constant feedback and interaction, even if that interaction is through a curtain of technology.

In Chapter 4, Dr. Xian discusses how distributed computing and constant monitoring can improve patient care, and in Chapter 3, Dr. Wang-Roveda gives a broad overview of how hospitals are integrating distributed computing into their daily operations [3, 2].

Another use of distributed computing is not to use it as an augmentation of a doctor or healthcare worker, but to use distributed computing as a medicine in of itself. One application is for Alzheimer's patients. Distributed computing can be used as essentially a vaccine against memory loss, helping people with dementia to remember basic tasks. Distributed computing also has uses for other types of mental disorders, for example, helping children with autism learn how to interact with the world.

Distributed computing can also be used as a diagnosis tool. By making consistent measurements, distributed computing can help to diagnose aspects that were impossible to quantify before, such as frailty in the elderly or the amount of eye contact made by children with autism. These uses of distributed computing are explained in greater depth in Chapters 3 and 6, this volume [2, 5].

In Chapter 4, this volume, we discuss a pervasive sensing platform that includes wireless "body sensor network" (BSN) and the "mobile base unit" (MBU) [3]. BSN acts as a sensor node and measures the physiological data using a group of sensors. Then, it is communicated using the MBU wirelessly for data transmission. It allows more portability. The patients could wear the sensors as watches, rings, clothes, etc.

Using mobile phones, we can study lifestyle and behavior that relate to the occurrence of a disease like smoking or overeating. They provide early evidence of an impending illness. We collect and analyze the data using mobile sensing. We present a mobile sensing toolkit called InCense and how it is used.

However, although sensors may generate a lot of new data, this data must first be processed into a useful form in order for this data to be useful. If distributed computing systems are going to be implemented in large scales in medicine, these systems must provide quality information that will make their use more efficient than face-to-face interactions. One of the main problems with things such as electronic health records at the moment is that they are not easy to use, are not standardized, do not give the doctors a good way of seeing a patient's medical history, and take up too much time to fill out. In order for distributed computing systems to be useful and for doctors to pay attention to them, they need to be robust and not overwhelm doctors with data. Readability is what makes data worthwhile. If a sick patient wearing a distributed system is allowed to go home, that distributed system should send the hospital an alert if the patient has a genuine emergency. However, if the patient does not have an emergency, the system should not send a false alert, which would take away resources from legitimate problems. This is discussed more clearly in Chapter 5, this volume [4].

Two of the biggest technical challenges in implementing distributed computing in medical applications are providing consistent power to the devices being used and determining the priority among events sensed by a network of devices or body sensors. These two factors help conserve the two main resources for any system—the energy of the machines in the system and the energy and time of the people in the system.

Wireless power transfer has recently become a very important topic. We mostly want longer battery life. There are three approaches that are discussed in this chapter: electromagnetic wave approach, inductive coupling, and magnetic resonance coupling. Wireless power transfer is also used in implantable medical devices. It is explained in more detail in Chapter 7, this volume [6].

The other method used is wireless body area networks (WBANs), which is a class of sensors that support a variety of health applications. It comprises a set of sensors and an energy-constrained fusion center. The objective of this chapter is to maximize the lifetime of this unique sensor network while ensuring proper physical activity. It is discussed in more detail in Chapter 8, this volume [7].

One of the methods used to determine whether an event has occurred or not is through Markov chains, or a MDP. MDPs, or partially observable MDPs, are used frequently in robotics systems for things such as path planning and map building. These algorithms are robust because they merely require that the robot be able to hold a belief in a particular state and can operate with limited knowledge of the outside world. In this case, the techniques enable distributed sensor nodes to adapt their energy output by changing their sampling frequency, and can take more samples when they believe something interesting is happening, and take fewer samples when the environment seems more routine. This algorithm and its implementation are described in detail in Chapter 9, this volume [8].

So far, we have talked about mostly implementing current technologies in medicine. Although some of the techniques are problems that have not yet been solved, the technology used is somewhat mature—the Internet already exists, and mobile phones already use accelerometers and other built-in sensors and can connect to other devices. In Section 1.6 we introduce new discoveries in materials and devices that will extend the reach of wireless computing.

1.6 NANOTECHNOLOGY IN MEDICINE

Distributed systems can already take advantage of current medical implants and other data sources; however, with these new medical sensors, we could do much more.

Why is nanotechnology revolutionary to medicine? Maybe this is too obvious. A better question may be the following: *How* will nanotechnology change medicine? It could potentially provide scientists with a new way to look at and affect the world. Ideas are important. The germ theory of medicine did not in of itself cure any diseases, but the idea of the germ theory of medicine allowed doctors and scientists to develop and use cures and prevention techniques because now they understood why certain illnesses were contracted and spread. It was a technique for how to understand the world. The discovery of DNA had a similar effect, because now you could understand why some people were more or less susceptible to certain disorders, etc. Nanotechnology is a broad field, but at its core, the idea of nanotechnology is to understand our world at a level that we don't understand it at right now. Nanotechnology will allow us to observe and change things that before were considered beyond our ability. It is transforming what was previously magic into science.

Applications of nanotechnology in medicine include but are not exclusive to the following:

1. Macro-sized materials with new properties because of the nanostructures that make up the material.
2. Nanoparticles as detection mechanisms.
3. Nanoparticles as drug delivery mechanisms.
4. Nanotechnology as implantable/injectable sensors for health monitoring.

Many uses of nanotechnology are special in large part because nanotechnology functions on such a small scale. Nanorobotics, for example, is a field that desires mostly to replicate the functions of macro-sized robots in a tiny form. The desired functionalities of nanorobots—actuation, sensing, and communication—exist in macro form, but have yet to be translated into the nanoscale. Medical nanorobots (while also being scary horror film gray sludge material), which have the capabilities of their macro counterparts, could perform functions such as drug delivery, diagnosis, or continuous sensing of various health parameters.

Medical nanorobots have a lot of potential. They could possibly decrease toxicity of cancer treatments by offering a more targeted precision on the cellular level. However, there are a number of challenges to be faced before medical nanorobots are viable. Some of those challenges are technical in nature. We do not know how to build or produce nanorobots in any quantity. Some so-called organic nanorobots have been produced, using modified bacteria, for example, salmonella, but these lack the control and communication that are necessary to truly describe them as nanorobots in a true sense. Other challenges are ethical or social in nature. Invisible, tiny robots that can track a person's health are rightfully scary, and appropriate bounds on this technology should be put in place. It is discussed in more detail in Chapter 10, this volume [9].

Nanoparticles are useful both because they have properties different from their macrosized counterparts and because they are very small. For example, gold nanoparticles behave differently in carbon nanotubes. Nanoparticles that can be controlled by magnetic field are called magnetic nanoparticles (MNs). Using this property, we can send drugs attached to the MNs and direct these to the required location using external magnetic field. To combat heat issues, magnetoelectro nanoparticles were used instead. There are three processes that are transmission and targeting, drug release, and drug intake. It is explained in detail in Chapter 11, this volume [10].

One special case of nanotechnology is DNA. The size of a single tRNA is 7 nm, although the length of a DNA strand is much longer—a few centimeters of DNA per cell, when uncoiled. DNA studies are not only a part of nanotechnology but also go far beyond nanotechnology. One particular application for DNA is DNA computing.

DNA computation is made possible by encoding the basic components of computer logic—input, output, and logic gates—in DNA strand interactions guided by Watson–Crick base pairing. This property of DNA also enables the design and construction of self-assembling nanoscale structures with defined features at molecular resolution. Recently, DNA computing and DNA nanotechnology were integrated to create a new generation of molecular machines capable of sensing their environment, processing the data, and actuating in order to write into the environment a desired output. Since DNA is a biological molecule naturally interfacing with the biology, biochemistry, and genetics of living organisms, it now provides novel strategies for treating diseases. These features are discussed in Chapter 12, this volume [11].

DNA nanotechnology is a special instance of nanotechnology that uses biology-created nanostructures. Other fields of nanotechnology rely on human-created nanostructures that would not form in nature without human intervention.

Many applications of nanomedicine are based upon the properties of nanoparticles. One category of nanoparticles is nanoparticles made out of graphene. Graphene, a 2D material, is an allotrope of carbon. (An allotrope is a specific configuration of atoms, in this case carbon atoms; other allotropes include diamond and graphite.) Graphene can be wrapped into fullerenes (spherical carbon forms), carbon nanotubes, and other shapes. It is strong, light, and flexible. Graphene also has incredibly high electron mobility at room temperature, making it an excellent conductor of electricity.

Graphene exhibits characteristics unlike any other material, and these characteristics can be used to create novel sensors. Graphene's high surface-to-volume ratio, high electrical conductivity, mechanical strength, and chemical stability provide sensing advantages. One application is to detect biomarkers, specific molecules that the body produces when it is diseased. The applications of graphene are discussed in Chapter 13, this volume [12].

Nanotechnology includes things such as computer chips made out of nanostructures. In these cases, it is not so important that the created items are small, but rather the properties of the new material are different from their macro-sized counterparts, and these characteristics can be used to develop new things. For example, obsessive–compulsive disorder (OCD), schizophrenia, and Parkinson's are assumed to be incurable because the macrosized medicines used were not responding the way we require, but recently nanotechnology has changed the scenario. Neuromorphic circuits that model the disorders like OCD, schizophrenia, multiple sclerosis, and Parkinson's are used to demonstrate behavioral differences with respect to healthy neural networks. Most of the circuits employ nanotechnology. It is explained in detail in Chapter 14, this volume [13].

Nanomedicine can be exploited to synthesize nanocomponents, which can be linked to microscale transporters and agents. The delivery of therapeutic agents to the site of treatment by navigating through the shortest vascular route so as to avoid circulation of highly toxic agents is an example of how nanorobotics can enhance medical interventions. We further explain the exploitation of the phenomena at all scales. It is discussed in detail in Chapter 15, this volume [14].

Nanomedicine's transversality means that the absence of radical transformations in the nanomedicine field does not mean that it is a failure. It means that it is a roadmap as to what is yet to come. There are three important areas in the contemporary biomedicine: predictive, personalized, and regenerative medicine. It intensifies and builds on the already existing tendencies within biomedicine. It is explained in detail in Chapter 16, this volume [15].

1.7 ETHICS OF MEDICAL WIRELESS COMPUTING

Although these technologies are promising, the fact remains that we should implement them in a safe, ethical way in order to reap the benefits. Misuse of technology is an old story, and there are specific fears that must be allayed before any of these innovations become realities, much less transform the way that the healthcare system

works. To that aim, we discuss the ethics of wireless computing in medicine from the extent to which a person's medical data should be available for research to the power of nanotechnology. In discussing these problems, we desire to mitigate the unintended consequences of wireless computing in medicine and guide a conversation about how to implement these technologies in a way that maximizes the benefits while respecting reasonable ethical boundaries.

Application of the new information and communication technologies to medical diagnosis, record keeping, and treatment can revolutionize healthcare. However, this presents many questions about appropriate ethical behavior in regards to these systems. The vulnerability of computerized databases presents challenges often categorized in terms of patient privacy, but the range of issues is much broader. It is discussed in detail in Chapter 17, this volume [16].

Equally important is the accountability of healthcare professionals, whose behavior the databases also record, in the context of medical care. It raises the question of how doctors should be judged in light of this new data. If a patient refuses help, or if a surgery goes wrong, should the doctor be held accountable? If a doctor knows that every action is recorded, will his or her behavior change? The ethics of the medical profession traditionally focused exclusively on the doctor–patient relationship, but in the context of publically funded healthcare and the necessity for an extensive epidemiological research to determine the differential effectiveness of treatments, the entire society becomes involved.

The basis for establishing ethical principles for the new technologies then expands to include issues of social inequality, political doctrine, investment in the welfare of children versus the elderly, and innumerable disputes concerning the extent to which government should monitor and control the behavior of citizens. Such issues become acute when the problems in question involve mental health, addictions, and what sociologists call the undue medicalization of deviant behavior. In Chapter 19, William Sims Bainbridge catalogs many of the ways that the technologies of wireless computing are transformative and draws upon classical perspectives on morality from philosophy and social science to understand their ethical implications [18]. It is discussed in the previous chapter [17].

In Chapter 18, Clark Miller et al. explore some of the specific dilemmas, which are mentioned in the previous chapter. They look at three prominent visions that ubiquitous healthcare promotes: the *knowledge-empowered individual*, *routine health surveillance*, and *digital medicine*, and explore some of the paradigms and effects that each of these three visions of ubiquitous healthcare may produce.

1.8 PRIVACY IN WIRELESS COMPUTING

Privacy is one of the issues that have to be solved in order for wireless computing to be widely adopted. Conversely, the sharing of data could help improve healthcare immensely, by allowing doctors to have an easy access to a patient's complete medical history and also by sharing results of various studies and allowing researchers to mine this data for new results.

One of the issues receiving a lot of attention is Electronic Health Records (EHRs). In 2009, the United States enacted the Health Information Technology for Economic and Clinical Health Act (HITECH), which aimed at increasing the use of EHRs (http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/html/PLAW-111publ5.htm). However, at this time, the adoption of EHRs remains at less than 50% and only 20% of doctors report that the health records are fully functional. (Are More Doctors Adopting EHRs? Retrieved 31 March 2011.) The standardization of Electronic Health Records needs to happen in order for EHRs to be useful, and EHRs should be shared across hospitals in order to be useful. However, the sharing of patient's data should be done carefully. Confidentiality, accessibility, and security are major issues in the creation of e-healthcare systems. E-healthcare by virtue of existing takes data once only shared between a doctor and a patient, and shares it with the company that is operating the e-healthcare system. That data is also stored in servers that can be hacked. Additionally, there are legal gray zones, such as insurance companies who could require the patient's EHRs before granting insurance. These issues and more are discussed in Chapter 19, this volume [18].

Due to the fundamentally different ways of treating patients, we have completely new ethical and privacy issues. The US intellectual property protection system has difficulty dealing with nanomedicine claims due to its interdisciplinary nature. The current laws and regulations are ambiguous when it comes to nanomedicine, and they need to be updated. It is discussed in detail in Chapter 20, this volume [19].

1.9 CONCLUSION

Wireless computing is a broad, interdisciplinary field that draws both from traditional distributed computing and nanotechnology. Applications of wireless computing have a potentially transformative effect on medicine, creating both opportunities for benefiting society and ethical pitfalls. Opportunities in distributed computing include creating integrated systems for hospitals, extending the reach of the doctor's office into the home, and using computing itself as a treatment for Alzheimer's and other cognitive diseases. Simultaneously, progress in nanotechnology has created new ways of sensing the world around us, opening up new ways of diagnosing, treating, and monitoring disease. As these technologies are integrated into medicine, we should also be aware of their effect on our society, how they change the definition of disease and medicine itself, and how these methods open up new vulnerabilities in data security and privacy. With a complete view of the technology and societal facets of wireless medicine, we can hope to gain the greatest benefits that the technology promises while respecting the rights of the people who use it.

REFERENCES

[1] MM Eshaghian Wilner, Arushi Gupta, Shiva Navab, Alekhya Sai Nuduru Pati, Gaurav Sarkar, "Chapter 2: Nanocomputing and Cloud Computing", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[2] Linda Powers, Kui Ren, Janet Meiling Wang-Roveda, "Chapter 3: Pervasive Computing in Hospitals, Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[3] Xiaojun Xang, "Chapter 4: Diagnostic Improvements: Treatment and Care" , "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[4] Luis A. Castro, Edgar Chavez, Jesus Favela, Jessica Beltran-Marquez, Rene Navarro, Moises Perez, Marcela Rodriguez, Eduardo Quintana, "Chapter 5: Collaborative Opportunistic Sensing of Human Behavior with Mobile Phones", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[5] Franceli N. Cibrian, Lizbeth Escobedo, Jose Mercado, Monica Tentori, "Chapter 6: Pervasive Computing to Support People with Cognitive Disabilities", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[6] Zhuochen Ge, MM Eshaghian Wilner, Renjun Liu, "Chapter 7 :Wireless Power for Implantable Devices: A Technical Review", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[7] Daphney Stavroula Zois, Sangwon Lee, Murali Annavaram, Urbashi Mitra, "Chapter 8: Energy Efficient Physical Activity Detection in Wireless Body Area Networks", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[8] Shuping Liu, Anand Panangadan, Ashit Talukder, Cauligi S. Raghavendra, "Chapter 9: Markov Decision Process for Adaptive Control of Distributed Body Sensor Networks", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[9] Amber Bhargava, MM Eshaghian Wilner, Wan Lee, Mike Schlesinger, Abhishek Uppal, Janet Cheung, "Chapter 10: An Introduction to Nanomedicine", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[10] MM Eshaghian Wilner, Sakhrat Khizroev, Gaurav Sarkar, Rakesh Guduru, Umang Sharma, "Chapter 11: NanoMedicine Using Magneto-Electric Nanoparticles" , "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[11] Ido Bachelet, Noam Mamet, "Chapter 12: DNA Computation in Medicine", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[12] Farid Menaa, Sandeep Kumar Vashist, Adnane Adbelghani, Bouzid Menaa, "Chapter 13: Graphene Based Nanosystem for Detection of Proteonic Biomarkers", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[13] Saeid Barzegarjalali, Rebecca Lee, Alice Parker, Sukanya Patil, "Chapter 14: Modeling Brain Disorders in Silicon and Nanotechnologies", "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[14] Sylvain Martel, “Chapter 15: Linking Medical Robots to Pervasive Computing”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[15] José J. López, Mathieu Noury, “Chapter 16: Nanomedicine’s Transversality—Some Implications of the Nanomedical Paradigm”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[16] William Sims Bainbridge, “Chapter 17: Ethical Challenges of Ubiquitous Healthcare”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[17] Gaymon Bennett, Clark A. Miller, J. Benjamin Hurlbut, Heather M. Ross, “Chapter 18: The Ethics of Ubiquitous Computing in Healthcare”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[18] Fredrick Japhet Mtenzi, “Chapter 19: Privacy Protection of EHR in e-Healthcare Systems”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

[19] Ayush Chaudhary, Pujal Trivedi, MM Eshaghian Wilner, Arushi Gupta, Ben Shiroma, “Chapter 20: Ethical, Privacy, and Intellectual Property Issues in Nanomedicine”, “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, John Wiley & Sons, 2015.

2

NANOCOMPUTING AND CLOUD COMPUTING

T. Soren Craig[1], Mary Mehrnoosh Eshaghian-Wilner[1], Nikila Goli[1], Arushi Gupta[1], Shiva Navab[2], Alekhya Sai Nuduru Pati[1], Kodiak Ravicz[1], Gaurav Sarkar[1], and Ben Shiroma[1]*

[1]*EE-Systems, University of Southern California, Los Angeles, CA, USA*
[2]*Broadcom, Irvine, CA, USA*

2.1 INTRODUCTION

Despite being polar opposites, nanocomputing, the apex of the miniaturization of computers, and cloud computing, the geographically distributed platform of computing, both promise to revolutionize the fields of computing and medicine.

Nanocomputing is the study of devices, paradigms, and applications that surpass the domain of traditional microcomputers by using physical phenomena and objects measuring 100 nm or less [1]. The shrinking of traditional transistor design is reaching its fundamental limits, running into subthreshold current problems and problems in achieving greater transistor density. Born out of traditional computational methods' failure to satisfy demands for even smaller computers, nanocomputing uses many of the properties that materials only exhibit at the atomic scale like quantum effects.

Cloud computing is the outsourcing of processing power away from individual computers to larger, more powerful, and more efficient servers through information networks like the Internet. This moves data storage and computing from PCs and laptops to large data centers, usually over the Internet, in order to save the individual the cost of more powerful hardware or larger storage capacity. The idea of distributive

*Authors are listed alphabetically by last name, as opposed to by order of contribution.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

computing is by no means a new one and has evolved from parallel computing, to distributed computing, to grid computing, and to now cloud computing.

Nanocomputing and cloud computing are two rapidly advancing fields that are changing traditional computing paradigms and offer solutions to the fundamental limits those paradigms are quickly approaching. These two technologies represent the two extreme ends of the spectrum. One is miniature and invisible to see, and the other can span the globe. Given their size and distributive nature, respectively, it is easy to envision a multitude of uses where the unique qualities of these two types of computing would make them far superior to traditional means. One such use in which nanocomputing and cloud computing both offer particular promise is its application within the field of medicine. The overview of nanocomputing and cloud computing in this chapter is meant to establish a baseline knowledge of these two innovative forms of computing, which the rest of this book will discuss in further detail.

This chapter is divided into two major sections. Section 2.1 talks about nanocomputing and its application in the medical field, and Section 2.2 describes cloud computing. Possible medical applications of nanocomputing are looked at with particular attention through the example of the use of spin-wave architecture to sequence DNA, which was chosen due to this author's familiarity with the topic. Both these sections serve as primers for the chapters to follow that talk about research and experiments in the field of nanomedicine and health applications of distributive computing, in which nanocomputing and cloud computing are integral parts, respectively.

2.2 NANOCOMPUTING

In this section, we discuss one extreme of computing: nanocomputing. See Sections 2.2.1, 2.2.2, and 2.2.3.

2.2.1 Overview of Nanocomputing Paradigms and Architectures

There are many ways in which the properties exhibited by electrons at nano level can be harnessed to perform computational functions. The concept of digital computing is simple: everything needs to be eventually broken down into 1s and 0s, known as binary form. Anything that can be represented in two distinguishable and controllable stable forms can be a potential candidate for digital computing. At nano level, there are quite a few examples of computing processes and architectures that fit this criterion about which we talk in Section 2.2.1.

In this section, we discuss nanocomputing architectures and technologies like molecular switches, carbon nanotubes, resonant tunnel diodes (RTDs), DNA, and protein computing.

2.2.1.1 Molecular Computing and Switches Molecular computing seeks to build computational systems wherein individual or small collections of molecules, often at the nano level, serve as discrete device components or play a significant role in them. Typical molecules used in these molecular switches are orders of magnitude smaller than the state-of-the-art silicon MOSFET. Solid-state electronic devices based on

molecular switches have been proposed as the active units in both nonvolatile random access memory circuits and as the reconfigurable bits for a custom configurable logic-based computing machine [2]. Molecular switches control the flow of electrons using the relative position of extremely small mechanical parts. This central element is based on a molecular switch tunnel junction that can be electrically switched between high- and low-conductance states. Alternative implementations of molecular switches use current to affect the way light is absorbed by a specific molecule [1]. Mounting evidence, both experimental and theoretical, confirms the molecule's role in the devices' switching mechanism [2–14]. This can potentially lead to computing systems with much higher density and performance and lower power use and cost than those of current silicon MOS technologies. However, while it is small size is what makes molecular computing so powerful, working on such a scale poses unique difficulties of its own. First, even though molecules can be synthesized in large quantities relatively easily, arranging them in an accessible way is far more difficult. Furthermore, it is equally difficult to ensure that every molecule stays in place. Second, although individual molecules have been shown to be switching, it is very difficult, if possible at all, to interconnect them or selectively interface them with microscale inputs and outputs.

2.2.1.2 Carbon Nanotubes As their name implies, carbon nanotubes are simply extremely small tubes that are comprised of pure carbon. Pure carbon takes two unique forms: (i) diamond, which is a three-dimensional lattice usually formed under high amounts of pressure and (ii) graphite, which easily separates into two-dimensional lattice sheets [1]. These sheets, when made sufficiently thin (1 atom thick) are referred to as graphene. The carbon nanotubes of most interest to nanocomputing are created when these graphene sheets form cylinders or single-walled carbon nanotubes. These carbon nanotubes have several novel properties and serve as the fundamental building block of many nanocomputing systems.

Carbon nanotubes' primary useful property is their high conductivity and corresponding low resistivity. These attributes allow carbon nanotubes to serve as ballistic conductors for electrons [1]. As a ballistic conductor, these nanotubes allow electrons to actually travel through their center along the axis with little resistance [1]. This makes them vastly superior in both size and efficiency to traditional wire conductors, which have resistances that increase inversely with their radii. However, nanotubes' utility to nanocomputing does not end with their superior conductive properties, but extends to all areas; carbon nanotubes have been successfully utilized to make switches, supports, and wires for nanoscale devices [1].

2.2.1.3 Resonant Tunnel Diodes (RTDs) RTDs take advantage of the quantum effects that affect electrons when dealing with them as individual particles on the nano level. Electron tunneling, which is a major problem for traditional transistors as they are scaled down, is now taken advantage of and the discrete energy states at which it is most likely to occur are used to parallel the discrete 1s and 0s computing requires.

Consider the problem of a "particle in a box." By this, we mean a particle, say an electron, confined to a small region in space by a potential energy distribution. This represents what is called an infinite potential "well"; namely, a region where

the electron is trapped by two barriers on the sides. The fundamental equation that governs the behavior of quantum mechanical particles, such as electrons, is the Schrodinger equation. The energy levels that an electron is allowed to have in this one-dimensional potential well can be easily obtained by an analytical solution of the Schrodinger equation. The result is as follows:

$$E_n = \frac{h^2 \pi^2 N^2}{2md^2}, \quad N = 1, 2, 3, \ldots$$

Here "h" is Planck's constant and d is the width of the well. In other words, the electron in the well cannot have just any energy, but must take one of the discrete values given by the above formula. In general, such "quantization" of energy levels also happens in the 3D case. This idea is at the heart of some of the quantum devices that we will discuss in this section. Now consider a region in space where a potential well is connected to two metal electrodes through barriers with finite heights and widths on the sides. This is an RTD. An electron can enter this region from outside, leave the region by overcoming the barrier heights (by, for instance, acquiring thermal energy and going to higher-energy levels), or move through the barriers by a process called quantum mechanical tunneling. What is interesting is that in the transport characteristics of this device, the effect of these discrete energy levels becomes completely visible.

Now imagine a voltage bias is applied to the structure that leads to a relative shift in the chemical potentials of the two contact electrodes. An electron will be able to tunnel through the device from one side to the other only if the biases on the two sides are such that there is an energy level in the well in the range where electrons exist on the left and empty states exist on the right, that is, when there is a level lower than $m1$ but higher than $m2$. (Remember that in the electrodes, all energy states up to the chemical potentials are filled with electrons.) Thus, as the applied bias is increased, every time a new energy level enters the range between the two side chemical potentials, there will be a peak in the device's current versus voltage curve.

If a "gate" electrode is placed below the device to enable us to move the energy levels up and down, then we can use this gate to control which level lies between the side chemical potentials, and therefore we can control the conductance of the device. This is the basis of a three-terminal switching device or resonant tunneling transistor.

Note that the conductance through this device cannot be modeled by simply considering two single barriers in series. In fact, what is essential here is the wave nature of the electrons and the resonance phenomenon in the well that leads to a high transmission probability of electrons from one side to the other, giving rise to the peaks in conductance. This is analogous to the transmission of light through a multilayer structure with layer thicknesses on the order of the wavelength of the incident light or smaller. Resonance phenomena there can lead to high transmission for a given set of layer properties (thicknesses and refractive indices) and incident wavelength. In general, this is the problem of a resonant cavity, which in the case of the RTD is a cavity for electrons.

2.2.1.4 Single-Electron Transistors As their name implies, single-electron transistors (SETs) realize the device functionality of a transistor by controlling the movement of an SET. Emulating the operation of a traditional transistor, the SET emits an electron to a small silicon island coupled to two external reservoirs (source and drain) through a tunneling barrier, and the potential barrier of the island can be controlled by a gate or multiple gates. Because electrons are dealt with in discrete quantities as individuals rather than as a current, an important effect to consider in the operation of SETs is the coulomb blockade. Given that like charges repel each other, when the gate is negatively charged, a situation when an electron is already present, other electrons are repelled, and thus becoming harder for additional charges to enter. Based on this effect and this functionality, several circuit applications in logic and memory have been proposed and simulated [15–20]. A review of single-electron transistor devices can be found in Ref. [21]. SETs can also be operated under an alternating voltage, and the frequency and amplitude of this signal is such that only one electron is transferred per cycle through the device. Such a device is called a single-electron turnstile [22].

2.2.1.5 Quantum Cellular Automata The quantum tunneling effect that enables intercellular interactions is used to develop cells known as quantum dots and electrons can "tunnel" through the barrier between the dots. The quantum cellular automata (QCA) has been extensively studied by a group of researchers at the University of Notre Dame for several years [23–25]. The basic idea behind QCA is that when the level of integration is very small, then cells interact with each other through quantum effects and tunneling. Utilizing quantum dots, the size of an elementary cell can be shrunk down to hundreds or tens of nanometers, and the intercell interaction can be realized via quantum tunneling without wires. Using this concept, simple cells have been developed mainly using five quantum dots called a quantum dot molecule. The five dots are close enough to enable electrons to tunnel between the dots. The barriers between cells are assumed to be sufficient to completely suppress intercellular tunneling. Two electrons occupy each cell. The occupancy can be stabilized because of the large energy splitting between different charge states of the cell. The Coulomb interaction between electrons in a cell acts to produce two distinct cell states with different charge configurations. If the barriers to tunneling are sufficiently high, the two-electron ground-state wave function in the cell will localize the two electrons on antipodal sites. This localization is due to Coulomb exclusion, a phenomenon closely related to the well-known Coulomb blockade of current and results in nearly exact quantization of charge in each dot. One major benefit of the QCA or electrostatics approach is that energy consumption is near negligible because only a few or fractions of an electron are used for each bit of information. Unfortunately, molecular QCA is based on the interconnection of individual molecules. While implementations of small QCA circuits have been reported, none are based on molecular quantum dots, due to our inability to position and interconnect individual molecules.

2.2.1.6 Nanowire Crossbar Arrays The basic crossbar array employs two overlaid but orthogonal arrays of parallel nanowires. A molecular monolayer can be formed between the two arrays, particularly at the crosspoint of two orthogonal

nanowires. Semiconductor nanowires, usually silicon nanowires and their variations, metallic nanowires [22, 26], and single-walled carbon nanotubes (SWNTs) [21] have been the most popular choices for the nanowires. The crossbar array architecture addresses the integration challenges by using a regular structure that seamlessly integrates devices and interconnects; by using the nanometer diameter of nanowires to form nanoscale crosspoints, which can host nanoscale molecular bundles; and by allowing independent choices of interconnects and devices, especially the flexibility in the choice of devices. Moreover, the large number of identical nanowires and crosspoints provides the redundancy for critical fault tolerance. The flexibility of the basic crossbar array architecture has led to numerous variations of crossbar array architectures, depending on the choices of nanowires, devices in the crosspoints, and addressing methodologies. It is important to note that the crossbar array architecture was first employed by MOSFET programmable array logic (PAL) [27]. An FET can also be formed at the crosspoint by doping the silicon nanowire [28]; many circuits have been designed or fabricated based on such FET crosspoints [29].

2.2.1.7 DNA and Protein Computing DNA- and protein-based computing is a form of nanocomputing, not to be confused with nanocomputing methods to sequence DNA, which uses the fundamental bonding properties of DNA to create sequences of DNA or proteins that can represent computational information. To understand this complex form of bioinspired nanocomputing, one must first recall the properties of DNA. DNA is a macromolecule that comprises trings of four other organic compounds: adenine (A), cytosine (C), guanine (G), and thymine (T) [1]. Furthermore, A only can bond to T, and G only to C [1]. This property allows for the creation of two discrete amino acid combinations that can be used to represent the high and low states necessary for computing. DNA computing relies on exploiting this natural property to solve each discretized part of large, complicated problems simultaneously [1]. Once a problem can be modeled with the DNA, the physical laws, which govern its bonding behavior, force it to come to the most stable energy state, which is modeled to be the solution to the problem [1]. This is called the principle of "constraint satisfaction." Constraint satisfaction flips the traditional computing paradigm and instead of modeling abstractions with physical phenomena, constraint satisfaction relies on the enforcement of abstract rules to bring physical phenomena into an order [1]. This allows for the computation of extremely difficult computing problems, which would be nearly impossible to solve using traditional methods. One such example is the solving of the Hamiltonian path problem with DNA computing.

Related protein computing also relies on constraint satisfaction to produce a computing mechanism. Like DNA, proteins are complex amino acids [1]. The proteins of interest here have properties that are similar to the bonding behavior of DNA in that they follow certain unchanging rules. For example, the way certain proteins fold or unfold can be exploited to represent two states [1]. However, this behavior is far more complex than DNA's bonding characteristics, and the added utility of computing with proteins is dubious. Thus protein computing has so far only been simulated.

2.2.1.8 Spintronics In spintronics, the spin of the electrons is controlled by various methods like applying a magnetic field, or using ferromagnetic objects. The orientation of electron spins represents the data stored, for example, if it is in the same direction as the applied magnetic field, it represents one form of data (either a "0" or a "1") and if it is opposite, it represents the other form of data. This is a new approach to electronics, where the information is carried out by the spin of the carrier, in addition to the charge [30, 31]. The performance of these devices is based on the carrier transport dependence on the spin of the carrier. In turn, carrier spin polarization can be controlled via different physical effects as spin–orbit interaction, spin injection from ferromagnetic contacts, and spin interaction with external magnetic field. All spintronic architectures described thus far operate according to the common scheme: (i) information is stored into spins as a spin orientation (along with or opposite to the external magnetic field), (ii) the spins, being attached to carriers, transfer information from one spin-based device to another through a conducting wire, (iii) spin polarization of the transmitted carriers effects the conductance of the recipient device, and (iv) information is read at the output terminal. Although the performance of the spin-based devices might be advantageous, the use of charge transfer for information exchange between the devices will significantly limit the performance of spintronic architectures, as it preserves the shortcomings inherent to the CMOS-based architecture: interconnection problem and RC delay for signal propagation. The downside of using spin wave is the size limitation due to the 50 μm attenuation range of spin waves.

Having discussed about various architectures on nanocomputing, in Section 2.2.2 we focus on a particular set of architecture on which we have worked.

2.2.2 Spin-Wave Architectures

In this section, we present a discussion about spin-wave architectures. The phenomenon of spin waves depends on the precession of the spin of electrons. The angle of spin determines the magnitude; a change in the angle of the spin induces a change in the spin of the next electron, and thus the spin propagates in the form of waves. This change in magnetic field induces a measurable amount of voltage. The speed of propagation and the frequency of operation achievable using spin wave-based systems are far greater than current technologies.

A thorough study of three nanoscale spin-wave architectures, the reconfigurable mesh, crossbar, and fully interconnected cluster leads to the discovery of several of their unique features. One of their significant characteristics is that both communication and computation are performed using spin waves. This introduces a new form of computing, so-called wave-based computing, which can be done in both analog and digital modes. Some of the features that make the nanoscale spin-wave modules highly parallel architectures are their concurrent write feature and their capability of transmitting multiple waves at the same time. In addition, they also allow multiple pairs of processors to intercommunicate over the same bus simultaneously.

Spin wave is a collective oscillation of spins in an ordered spin lattice around the direction of magnetization. The phenomenon is similar to the lattice vibration, where

atoms perform oscillation around its equilibrium position. The magnitude of the spin wave is determined by precession angle. A propagating spin wave changes the local polarization of spins in the ferromagnetic material. In turn, the change in magnetic field results in an inductive voltage. Recent published experimental results indicate that an inductive voltage signal of the order of microvolts produced by spin waves is detectable at the distances of up to 50 μm at room temperature. Typically, the speed of the spin wave is 10^4 m/s. Based on this phenomenon several logic devices have been created, some of them are described in Section 2.2.2.1, after that we see how interconnection models based on spin wave can be realized.

2.2.2.1 Spin Wave-Based Logic Devices Spin waves can be used to provide an "LC" coupling of devices, without dissipative resistance. With the spin-wave concept, the spin rotates as a propagating wave and there is no particle (electron/hole) transport. The core of the spin-wave bus demonstration structure consists of a ferromagnetic film (e.g., NiFe) grown on a semi-insulating substrate. The film is polarized along the *X*-axis. There are three asymmetric coplanar strip (ACPS) transmission lines on the top of the structure. The lines and the ferromagnetic layer are isolated by the silicon oxide layer. The thickness of the ferromagnetic layer can be as thin as tens of nanometers, and the thickness of the oxide layer is several hundreds of nanometers. The dimensions of the ACPS lines are defined by the frequency of the transmitting signal. Each of the ACPS lines can be used for spin-wave excitation and detection. A voltage pulse applied to the ACPS line produces magnetic field perpendicular to the polarization of the ferromagnetic film, and, thus, generates a spin wave (spin-wave packet). Being excited, spin wave propagates through the ferromagnetic film. As it reaches the nearest ACPS line, the amplitude and the phase of the spin wave are detected by the inductive voltage measurements. For example, the edge ACPS lines can be considered as the input ports, and the ACPS in the middle as the output port. The middle ACPS line detects the inductive voltage produced by the *superposition* of two waves. Depending on the relative phase of the spin waves, the amplitude of the inductive voltage may be enhanced (when two waves are in phase) or decreased (when two waves are out of phase) in comparison to the inductive voltage produced by a single spin wave.

The spin-wave architectures can be used to achieve various logic functions like AND, NOR, etc. on the same device. This is possible by controlling the initial phases of the spin wave and the voltage threshold to which the output voltage is compared.

Khitun and Wang showed how the two-bit gates—AND and OR—and the one-bit—NOT—can be implemented on the prototype three-terminal device. In this section, we first briefly explain how they have implemented these logic gates, and next we show how to extend their design to implement other logic functions such as NAND, NOR, XOR, and XNOR.

The input information is coded into the polarity of the voltage pulse applied to the input ACPS lines and is detected at the output port. The edge ACPS lines are considered as the input terminals, and the ACPS line shown in the middle is the output port. The output logic state is determined by comparing the detected voltage to a reference voltage, V_{ref}. Assume the measured inductive voltage produced by a

single spin wave at the output port is V_1. The reference voltage V_{ref} can be a function of V_1 as shown below. The measurement is performed at the moment when two spin-wave packets arrive to the detecting ACPS line t_m. The exact choice of t_m depends on the logic function one needs to realize.

The one-bit NOT gate can be obtained measuring the inductive voltage, V_{out}, at the output line produced by spin waves excited by the input line. Taking $t_m = \pi g/v_{ph}$ (where g is the distance between the contacts and v_{ph} is the spin-wave phase velocity), and $V_{ref} = 0$ V, we achieve the required logic correlation. At this t_m, the spin waves are detected with 180°-phase difference comparing to the original phase.

Next, the two-bit AND gate can be realized when $t_m = 2\pi g/v_{ph}$, and $V_{ref} = V_1$. The output voltage ranges between +27 and −27 mV, and input voltage varies between +1 and −1 mV when $V_{ref} = 0.5$ mV. Two spin-wave packets coming in phase enhance the amplitude of the produced inductive voltage and cancel each other when coming out of phase. The nonzero reference voltage is needed to avoid the effect caused by the finite size of the detecting ACPS line. The OR gate can be realized by analogy, by taking $t_m = 2\pi g/v_{ph}$, and $V_{ref} = -V_1$. The output voltage ranges between +27 and −27 mV, and input voltage varies between +1 and −1 mV when $V_{ref} = -0.5$ mV.

Extending the same idea, a two-bit NAND gate can be implemented when $t_m = \pi g/v_{ph}$ and $V_{ref} = -V_1$. $V_{ref} = -0.5$ mV. Similarly, a two-bit NOR gate can be realized when $t_m = \pi g/v_{ph}$ and $V_{ref} = V_1$. $V_{ref} = 0.5$ mV for NOR implementation.

To realize XOR and XNOR gates, we need two reference voltages, V_{ref1} and V_{ref2}. In these two gates, we define the output to be in state "1" if the detected output voltage is between V_{ref1} and V_{ref2} and in state "0" if it is greater than V_{ref1} or less than V_{ref2}. For both gates $V_{ref1} = V_1$ and $V_{ref2} = -V_1$. For an XOR gate $t_m = 2\pi g/v_{ph}$, while in an XNOR gate $t_m = \pi g/v_{ph}$. For both XOR and XNOR, $V_{ref1} = 0.5$ mV and $V_{ref2} = -0.5$ mV.

The significance of efficiently implementing standard combinational arithmetic and logic modules is due to the fact that they are the basic blocks of any digital system. All the sequential systems can be implemented by just adding latches to the combinational circuit. The universality of the combinational modules ensures the possibility of realizing any logic switching function. A set of combinational modules are said to be "universal" if any combinational system can be implemented using the elements just from that set. In addition to logic gate sets such as {AND, OR, NOT}, {AND, NOT}, {OR, NOT}, {NAND}, and {NOR}, standard combinational modules can generate universal sets as well. For instance, a network of multiplexers (MUX) can implement any switching function according to Shannon's decomposition [32]; therefore, the set {MUX} is universal. {Decoder, OR} is another example of a universal set.

2.2.2.2 Spin Wave-Based Interconnection Networks Now that we have looked at how the phase logic helps achieving basic logic functions, we now focus our attention to interconnection models. Interconnection models are useful for simulating the complex neural networks, which have highly complex interconnection networks. Spin-wave research has shown three architectures to be particularly viable, crossbar networks, reconfigurable meshes, and interconnected clusters.

Crossbar Networks In this architecture, the ferromagnetic spin-wave buses are interconnected in a configuration that is similar to a general crossbar formation. The architecture described here, while requiring the same number of switches as a standard crossbar, is capable of simultaneously transmitting N waves to each of the spin-wave paths. As compared to the known molecular crossbars, this design is fault tolerant because in case there is a failure in one of the N channels, any of the other channels can be reconfigured to transmit the data. This is possible since all the channels are accessible by all the ports and each channel can handle multiple data.

Crossbars are attractive architectures because they can realize any permutations of N inputs to N outputs. However, their main shortcoming is due to the fact that N^2 switches are used to transmit only N pairs of data. As mentioned earlier, each spin-wave bus is capable of carrying multiple waves at any point of time. Therefore, each of the N inputs can broadcast its data to all of the N outputs in parallel. In other words, using this type of architecture it would be possible to efficiently realize highly connected type of computations such as the Hopfield neural network model.

An important element required for the spin-wave crossbar is the spin-wave switch. We define a spin-wave switch as a device which has an externally controllable magnetic phase. In the "on" state, the switch transmits spin waves, while in the "off" state it reflects any incoming spin wave.

A ferromagnetic film is divided by a region of diluted magnetic semiconductor (DMS) used as a magnetic channel, in which magnetic phase is controlled by the applied electric field via the effect of hole-mediated ferromagnetism [33]. A metal–insulator–DMS structure was used for experimental study of the effect of hole-mediated ferromagnetism in Ref. [33]. A negative gate bias increases the hole concentration in the DMS region resulting in the paramagnetic-to-ferromagnetic transition, whereas a positive bias has an opposite effect. The ferromagnetic film is divided by the DMS cell. Spin waves can propagate through the DMS cell only if it is in the ferromagnetic phase. While in the paramagnetic phase, the DMS cell reflects all incoming spin waves.

As shown, there is a common p-doped DMS layer for the whole structure. The magnetic phase in every region is controlled by the hole concentration. Using conducting ferromagnetic films (e.g., NiFe), it is possible to control the hole distribution along the structure by the applied electric field. The hole concentration is high when both the top and the bottom ferromagnetic films are biased by a negative voltage. There is a self-alignment mechanism making the proposed structure tolerant to structure imperfections.

Reconfigurable Meshes A nanoscale spin-wave reconfigurable mesh of size N^2 consists of an $N \times N$ array of processors connected to a reconfigurable spin-wave bus grid, where each processor has a locally controllable bus switch [34].

In the proposed spin-wave reconfigurable mesh architecture with spin-wave buses, a set of column spin-wave buses on the bottom and a set of row spin-wave buses on the top are connected via the spin-wave switches. Each switch is placed at the grid point of the mesh. Basically, the nanoscale spin-wave reconfigurable mesh is similar to the standard reconfigurable mesh, except for the spin-wave buses and switches.

Each of the spin-based devices serves as a one-bit input/output port. One spin-wave packet excited by one port can be superposed with many packets generated by the nearby ports. The length of interaction among the spin-based devices is limited by the spin-wave attenuation caused mainly by the scattering on phonons. The length of interaction is defined by the material properties of the ferromagnetic film, film size, and the operation temperature and has been experimentally found to be as high as 10 μm.

Interconnected Cluster The architecture presented in this section is a nanoscale fully interconnected architecture [35]. In this architecture, each node can broadcast using spin waves to all other nodes simultaneously and, similarly, each node via spin waves can receive and process multiple data in parallel. The significance of this design is that the communication between the nodes can be done in constant time. This is a significant improvement considering the $\Omega(\log N)$ lower-bound on the time delay for implementing such networks in VLSI using traditional electrical interconnects. Another key advantage is that the spin-wave fully interconnected network can be laid out in $O(N^2)$ area as opposed to $O(N^4)$, and of course the unit of area is in order of nanometer as opposed to the standard micron technology. In this architecture, information is encoded into the phase of spin waves, and no charge is transmitted. As a result of this, power consumption can be significantly reduced in this architecture comparing to other nanoscale architectures.

The area requirement of this architecture is $O(N^2)$ as opposed to the $O(N^4)$ area requirement if electrical interconnects were to be used. We should also note that in this architecture all the distances are in "nanoscale." Also unlike the network with electrical interconnections, in which only one transmission can be done at a time, in this architecture multiple simultaneous permutations are possible. This can be done by transmitting the spin waves over different frequencies. The information is coded into the phase of the spin waves in the sender and is detected by the receivers. In addition, within each frequency, data can be sent to one or more other nodes from each node.

Normally, in architectures where the phases of the waves are the means of information transmission, the exact location of the nodes with respect to the size of topology is an important design issue. The distance between the sender and receiver has to be a multiple of the wavelength; otherwise, the receiver might receive the wave at a different phase. For instance, if the wave is received with a π radian phase shift, a "0" is received instead of a "1" or vice versa. In this design, however, this is not an issue since the wavelength of spin waves is considerably larger than the distance between the nodes. The speed of spin waves is around 10^5 m/s. Assuming the input frequency range of 1–10 GHz, the wavelength will be in the order from 10^{-4} to 10^{-5} m, while the distances are at the nanoscale or 10^{-9} m. In other words, the wavelengths of the spin waves are orders of magnitude greater than the distances between the nodes. Therefore, all the nodes receive the same phase regardless of their location, and there is no need to place the nodes at specific distance relative to the other ones.

Having discussed about spin-wave architecture, we will now proceed with the discussion on its medical applications in this section.

2.2.3 Medical Applications of Spin-Wave Architectures

Nanocomputing, in specific spin-wave architecture, finds many applications in the medical field. The applications of DNA sequencing and modeling neural networks are given here as illustrative examples of the capabilities of both spin-wave architectures and nanocomputing and hopefully inspire one's imagination to think about what other tasks these new computing paradigms might one day accomplish. In this section, we discuss about the two main applications: DNA sequencing and modeling neural networks.

2.2.3.1 DNA Sequencing A major task in DNA sequencing is multiple sequence alignment (MSA). It involves figuring out the best alignment in a set of sequences. This process has a very high time complexity. By using spin wave-based mechanisms and partial order MSA one can reduce the problem of complexity to a large extent.

The rapid accumulation of biological data has necessitated more robust databases and data-processing algorithms. Consequently, new areas of computational biology are being created rapidly, combining the biological and informational sciences. To illustrate the complexities of some of the problems in computational biology, let us consider the genetic material in all living organisms—DNA. DNA is a polymer of nucleotides, where each nucleotide contains one of four bases: A, G, T, or C. An average gene in a human genome has 30,000 base pairs. With 30,000 genes estimated in each human genome, there are roughly three billion base pairs. A challenging task in computational biology deals with the finding of the "best" alignment among a given set of sequences. This is called a MSA.

The alignment of two sequences of length L can be solved via dynamic programming in $O(L^2)$ time. Extending it to N sequences, it will take $O(L^N)$ time, and in some cases the MSA problem is NP-complete [36]. Recent work includes CLUSTLW, T-COFFEE, etc. In one of the pioneering papers by Chris Lee, he demonstrated the advantages of applying a graph theoretical approach to the MSA problem (partial-ordered (PO) alignment), and many graph-based techniques have been proposed since then as explained in Refs. [36–39]. Lee proposed the PO-MSA and demonstrated its application in pair-wise alignment as well as its application in progressive pair-wise alignment [37, 38].

Using the process of forming PO-MSA graph (PO-MSAG) as a case study, one can see how the spin wave-based mechanisms assist in this task. Assuming there are O(N) sequences of length $O(L)$ that are already aligned, it will take at least $O(NL)$ time to simplify the sequences sequentially into a PO-MSAG. Here, spin wave-based mechanisms present a set of fast and parallel algorithms to generate a PO-MSAG for any given set of aligned sequences at the hardware architecture level. The hardware can then be fabricated as a low-cost chip that can be used as a coprocessor with a more powerful processor.

Two reconfigurable mesh architectures will be used to implement the solution: the nanoscale spin wave and the microscale electrical VLSI. The problem is first separated into two cases: data sequences with a fixed amount of data variations and those with an arbitrary amount up to $O(N)$. In the fixed variations case, the spin wave

and the VLSI architectures can construct the PO-MSAG in $O(1)$ time. As for arbitrary $O(N)$ variations, the spin-wave architecture will construct the PO-MSAG in $O(1)$ time, whereas the VLSI one takes $O(N)$ time.

The run times of both architectures are the same, $O(1)$. However for an arbitrary number of variables, the spin-wave architecture will have an $O(1)$ time complexity as opposed to an $O(N)$ time complexity using the other version.

2.2.3.2 Modeling Neural Networks This is the second important medical application of spin wave. The spin-wave architecture can also be used to model the complex interconnections of the neural networks to simulate how the neurons communicate. During the past several decades, researchers have developed electronic models of neurons designed to emulate neural behavior with electrical signals that mimic in some ways the measured potentials of biological neurons. However, due to certain inherent limits in VLSI technology, emulation of complex neural network models has not been feasible.

There are several challenges in realizing a neural network model. One of the most difficult of the challenges is the massive interconnectivity. Neurons in the human cortex possess on the average of 10,000 postsynaptic terminals, amounting to massive fan-in and fan-out [40]. Also, multiple synapses can converge (fan-in) to a single postsynaptic terminal, either from a single oversized presynaptic terminal or from multiple presynaptic terminals. The structure of the dendritic tree can be quite complex as well, affecting the processing of the neuron [41]. In this chapter, we are focusing on solutions to the interconnection challenges. Fortunately, many neural connections are local, and our solution exploits this property.

A suitable topological candidate for emulating neural circuits would be a fully interconnected network, where each of the N computing nodes could send (or receive) data to (or from) all the other nodes. However, as explained earlier, implementing such a network on a CMOS VLSI chip is not practical due to inherent interconnection limits in this technology. For instance, the VLSI area of a fully interconnected network of multiprocessors on a chip is $O(N^4)$, which is unreasonably large. Furthermore, the implementation of such an organization would require having nodes with $O(N)$ fan-in and fan-out, which also is not practical in VLSI. If constant degree nodes were to be used instead, then there would be an $\Omega(\log N)$ intercommunication delay lower-bound in realizing such a network in VLSI using electrical interconnects. Recent publications [42] discuss the feasibility of modeling neural networks using future nanoscale CMOS technology. The replacement of electrical wires with free-space optical interconnects could significantly improve the VLSI interconnection area and fan-in fan-out limitations.

Using spin waves, each neuron can broadcast to all other neurons simultaneously, and similarly a neuron can concurrently receive and process multiple data. Therefore, in these architectures, the total weighted sum to each neuron can be computed by the sum of the values from all the incoming waves to that neuron. In addition, using the superposition property of waves, this computation can be done in $O(1)$ time, and neurons can update their states quite rapidly. In the following, we describe the Hopfield model and show how it can be realized on spin-wave architectures.

The Hopfield model, an associative or content-addressable memory model, was proposed by John Hopfield. The publication of his work significantly contributed to the renewed interest in research in artificial neural networks [43]. In this model, each processing device (neuron) i has two states: $\alpha_i = 0$ ("not firing") and $\alpha_i = 1$ ("firing at maximum rate"). When neuron i has a connection made to it from neuron j, the strength of connection or so-called weight of this connection is defined as w_{ij}. ($w_{ij} = 0$ for nonconnected neurons). The input coming from neuron j to neuron i is shown as I_{ij}. The total input to neuron i, I_i, is computed as the weighted sum of all its inputs. For each neuron i, there is a fixed threshold θ_i. Each neuron randomly and asynchronously evaluates whether its total input is above or below threshold and readjusts accordingly [44, 45].

$$I_i = \sum_j I_{ij} * w_{ij} \quad \alpha_i = \begin{cases} 1, & I_i \geq \theta_i \\ 0, & \text{otherwise} \end{cases}$$

Section 2.3 talks about the second half of the chapter, which deals with cloud computing that essentially holds the key to pervasive and ubiquitous computing used extensively in the medical field.

2.3 CLOUD COMPUTING

Nanocomputing is discussed in Section 2.1, and we discuss the other extreme, cloud computing, in Section 1.3. See Sections 2.3.1, 2.3.2, and 2.3.3.

2.3.1 Description and Evolution of Cloud Computing

In this section, we first talk about how cloud computing evolved. Modern cloud computing started in the 1990s, but its roots can be traced back to the 1960s when John McCarthy instigated the time-sharing development system and predicted that computing will be provided as a public utility.

Cloud computing evolved from basic IT resources, which had simplistic and homogenous resources. Homogenous computing uses one or more machines of the same type and has provided adequate performance for many applications in the past. However, as its use increased to technical computing, the end users require high-performance computing, which cannot be provided by homogenous set of commodity hardware. New technologies came up, which provided much improvement in power density and efficiency. Heterogeneous computing (HC) was one such evolved technique. HC is a special form of parallel and distributed computing that could be applied to different levels of computing organizations. An HC system includes heterogeneous machines, high speed network, interfaces, operating system, communication protocols, and programming environments.

One other important milestone in the process of evolution of cloud computing is grid computing; computations are done using geographically distributed set of connected autonomous computers. Grid computing provides a virtual pool of

computation resources, but it's different from cloud computing. It specifically refers to leveraging several computers in parallel to solve a particular individual problem or run a specific application. On the other hand, cloud computing refers to leveraging multiple resources to deliver a unified service to the end user. Cloud computing provides tools and techniques to compute intensive parallel applications with lesser expenditure compared to traditional parallel computing techniques. Cloud computing shares its characteristics with certain models such as the following: client-server, grid computing, mainframe computing, utility computing, and peer-to-peer.

Having discussed about the evolution of cloud computing, in Section 2.3.1.2, we discuss about the description and features of cloud computing. However, before that, it is important to stress on the mapping technique used for heterogeneous computation, which now is also applicable to distributed models that led to the evolution of cloud computing. Cluster-M is one of the earliest models and is still the fastest known algorithm for mapping arbitrary task graphs onto arbitrary system graphs. In Section 2.3.1.1, we will give an overview of Cluster-M.

2.3.1.1 Cluster-M Cluster-M is a programming tool that facilitates the design and mapping of portable parallel programs. Cluster-M has three main components: specification module, representation module, and mapping module. In specification module, machine-independent algorithms are specified and coded using program composition notation (PCN) programming language. Cluster-M specifications are represented in the form of multilayered clustered task graph called Spec graph. Each clustering layer in Spec graph represents a set of concurrent computations called Spec clusters. A spec graph can also be obtained by applying one of the appropriate Cluster-M algorithms to any given task graph. A cluster-M representation represents a multilayer partitioning of a system graph called Rep graph. At every partitioning layer of Rep graph, there are number of clusters called Rep clusters. Each Rep cluster represents a set of processors with a certain degree of connectivity. The clustering is done only once for a given task (system) graph independent of any system (task) graphs. It is a machine-independent (application-independent) clustering; therefore, it is not necessary to be repeated for different mappings. For this reason, the time complexities for the clustering algorithms are not included in any time complexity of cluster-M mapping algorithm.

Basic Concepts There are a number of reasons and benefits in clustering tasks and system graphs in the cluster-M fashion. Basically, Cluster-M clustering causes both task and system graphs to be partitioned so that the complexity of mapping problem is simplified and efficient mapping results can be obtained. In clustering an undirected graph, completely connected nodes are grouped together to form a set of clusters. Clusters are then grouped together again if they are completely connected. This is continued until no more clustering is possible. When an undirected graph is a task graph, then doing this clustering essentially identifies and groups communication-intensive set of task nodes into number of clusters called Spec clusters. Similarly for a system graph, doing the clustering identifies well-connected set of processors into number of clusters called Rep clusters. In the mapping process, each of the

communication-intensive set of task nodes (Spec clusters) is to be mapped onto communication-efficient subsystem (Rep cluster) of suitable size. Note that mapping of undirected task graphs onto undirected system graphs is referred to as the allocation problem. In Bokhari's algorithm [46] where the time complexity is $O(N^3)$, it is assumed that $M = N$, where M and N are the number of nodes in the task and system graphs, respectively. By making the same assumption, Cluster-M algorithm becomes $O(N^2)$ and is $O(N)$ times faster. However, when adding the time complexity of Cluster-M clustering algorithms ($O(N^2)$ for clustering uniform-directed graphs (task graphs of size N) and $O(N^3)$ for clustering undirected graphs (system graphs of size N) to the cluster-M mapping time complexity, it leads to total time complexity of $O(N^3)$, which is still the same as Bokhari's. Please note that Cluster-M clustering algorithms are done only once for each task graph, and once for each system graph, and are not repeated for every new pair of task and system graphs that are to be mapped.

Clustering directed graphs (i.e., directed task graphs) produces two types of graph partitioning: horizontal and vertical. Horizontal partitioning is obtained because, as part of clustering, we divide the directed graph into layered graph such that each layer consists of a number of computation nodes that can be executed in parallel and a number of communication edges incoming to these nodes. The layers are executed one at a time, hence mapping is done one layer at a time. This significantly reduces the complexity of mapping problem since the entire task graph need not be matched against the entire graph.

Vertical graph partitioning is obtained because as part of clustering the nodes from consecutive layers are merged or embedded. All the nodes in a layer are merged to form a cluster if they have a common parent node in the layer above or a common child node in the layer below. Doing this traces the flow of data. Both vertical graph partitioning and horizontal graph partitioning are performed by the clustering in a bottom-up fashion. Cluster-M mapping will then be performed in top-down fashion by mapping the Spec clusters one layer at a time onto the Rep clusters.

With this description of Cluster-M and brief description of earlier technologies that led to the evolution of cloud computing, we now present the description and features of cloud computing in Section 2.3.1.2.

2.3.1.2 Description and Features of Cloud Computing In the simplest terms, cloud computing means storing and accessing data and programs over the Internet instead of the computer's hard drive. Cloud computing introduces a new Internet-based environment for on-demand dynamic provision of reconfigurable computing resources. The definition of cloud computing provided by the National Institute of Standards and Technology (NIST) states that "cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage) that can be rapidly provisioned and released with minimal management effort or service provider interaction." The main motive behind the development of cloud computing is to allow complex computation remotely using the powerful backend servers, thereby increasing the throughput of the utility and making powerful resources more accessible to

the public. With growing network, these shared resources can be provided as services for extended computation applications and also for cell phones or devices with limited computation power. The advent of cloud computing has started the trend of the move of data from local devices like personal computers and mobile devices to data centers. As a more practical aside, cloud computing also keeps the client away from the hassle of setting the configuration of the system, which is responsible for managing the data, computing, and providing access to storage services, and leaves maintenance and upkeep to trained professionals.

Before beginning with actual architecture of the cloud computing system, it is important that we glance through the highlights of this technology. Cloud computing has three main areas where it is superior to traditional computing—ease of use, cost, and range of capabilities.

Cloud computing is extremely easy for users to access. Usually the only thing that is required for one to take advantage of some form of cloud computing is an Internet connection and a browser. At that point, the user can access a wide array of powerful computing capabilities like server time and network storage without the need of human assistance. Furthermore, because the cloud is accessed from one's personal computer or some sort of terminal connection, the interaction with the cloud is usually familiar in terms of user interface regardless of the device used and service provided. This back-loading of computational resources allows users to use a broader range of devices as they are not constrained by computational requirements and thus encourages the use of diverse and heterogeneous client platforms.

By moving computing power away from the end user, cloud computing ultimately greatly reduces the cost of computing. Cloud computing providers use large-scale servers and other data-storage devices, which become more cost effective as their size increases. Furthermore, because providers allow many users to utilize their services, their server space and computing power is far more efficiently used. Additionally, with storage and even applications running off the cloud, users no longer need to take care of maintenance and upkeep, a task better left to trained professionals. Finally, the utilization of cloud services can be easily measured whether that be in the amount of data stored or operations performed, and thus users can be appropriately and accurately charged.

Cloud computing also provides a much wider and deeper array of computing services to the user than they would otherwise have access to. An interesting application of this capability is that the users' ability to access the cloud from multiple devices allows them to use their devices as terminals to the main cloud—a service that could cost hundreds if not thousands of dollars if they were to establish this for themselves. Unexpectedly, cloud computing is also more reliable than traditional point-to-point modes of communication. These resources are also elastically provisioned insofar as they are allocated only where they are required, ensuring maximum efficiency and increasing their availability to other consumers.

Unfortunately, cloud computing distributive design opens it up to various security and privacy issues that do not affect traditional computing. These security and privacy concerns are the primary ones many users have with cloud computing and also the main factors in slowing down cloud computing's adoption by consumers.

Users store valuable personal as well as professional data on cloud servers, which could be used to deleterious means if obtained by the wrong people. Furthermore, there are concerns that cloud computing service providers themselves might misuse the user data. These security and privacy concerns are further elaborated upon with special regard to their implications for the medical field in later chapters. Thankfully, like traditional computers cloud computing systems can indeed be made safer when more resources are devoted to security—for example, providers could invest in additional firewalls or levels of encryptions in order to better safeguard the user data.

We will now present the structure of cloud computing in Section 2.3.2.

2.3.2 Structure of Cloud Computing Systems

The basic architecture of cloud computing systems can be divided into two major parts—the front end and the backend. Generally speaking, the front end of any system is what the user sees and interacts with. In the case of cloud computing the front end would consist of the user's device and applications that they use to gain admittance to and interact with the cloud. Conversely, the backend is comprised of the cloud servers and mainframe computers where the user's data is stored and operations carried out, respectively. The link between the front end and the backend of cloud networks is usually the Internet, which ultimately limits interaction between the front end and the backend with bandwidth restrictions.

Figure 2.1 provides a high-level view of cloud computing architecture going from the front end (top) to the backend (bottom). The client is the user who interacts with an application, typically a cloud user interface application or simply an Internet on their platform or device. This device is connected to the cloud over an Internet infrastructure, which consists of the fiber optic cables and satellites that make the Internet possible. Finally, the back-end servers are where the data is actually stored and operations are performed.

2.3.2.1 Service Models of Cloud Computing Service delivery in cloud computing comprises three different service models: infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS). The three service models or layers are completed by an end-user layer that encapsulates the end-user perspective on cloud services. If cloud users access services on the infrastructure layer, for instance, they can run their own applications on the resources of a cloud

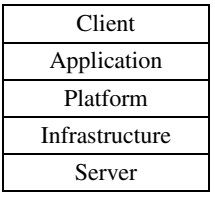


FIGURE 2.1 High-level conceptualization of cloud computing.

infrastructure and remain responsible for the support, maintenance, and security of these applications themselves. If they access a service on the application layer, these tasks are normally taken care of by the cloud service provider.

Software-as-a-Service By using SaaS systems, one can eliminate the need to install applications in the client system. Providing complete applications to a cloud's end user, this service is mainly accessed through a web portal and service-oriented architectures based on web service technologies. The important features of this include Internet-based services and management and maintenance of software from a centralized location with remote accessibility of the application by the users. Examples of this type of service providers are NetSuite, IBM, Oracle, Google, and Microsoft.

Platform-as-a-Service PaaS systems provide a computing platform, such as an infrastructure with all the applications required by the clients on it. This saves the user the trouble of buying licenses and installing the software. The principal users of this layer are developers seeking to develop and run a cloud application for a particular platform. They are supported by the platform operators with an open or proprietary language, a set of essential basic services to facilitate communication, monitoring, or service billing, and various other components. For instance, to facilitate startup or ensure an application's scalability and/or elasticity. Distributing the application to the underlying infrastructure is normally the responsibility of the cloud platform operator. This kind of service also helps developers to get a better understanding of the cloud environment, which in turn helps to maintain the software and makes updating it easier. Some major players in this area are Microsoft's Azure, GAE.

Infrastructure-as-a-Service IaaS service provides the infrastructure like hardware as a form of service. The client does not have to purchase the required data centers, servers, or other resources. The users are charged as per the duration for which the services are used. This reduces the cost of the services and ensures much faster delivery of services to the customers. Some companies that provide such services are Joyent, Flexiscale, and Rackspace.

2.3.2.2 Cloud Deployment Models With most organizations focusing on leveraging the cloud in order to cut capital expenditure and control operating costs, there is an aggressive growth in business for cloud adoption. However, the cloud can bring security risks and challenges for IT management, which can be more expensive for the organization to deal with, even considering the cost saving achieved by moving to the cloud. Therefore, it is very important for businesses to understand their requirements before opting for various deployment models available on the cloud. Now, the implementation of the cloud network itself can be of different types. Currently there are four types of implementations [26]: public, private, hybrid, and community cloud systems.

Public Cloud Public cloud allows indiscriminant access to everyone through an interface. The interface can be as simple as a web browser. Most of such services are "pay as you use," implying that users only have to pay for the time interval they use

the service. As mentioned before, this helps in improving the cost efficiency. On the downside, public cloud systems lack in adequate security, as the data stored can be hacked and accessed by anyone who has the technical know-how. By verifying the identity and validating both the service provider and the client, some level of security can be ensured, but it might prove to be inadequate in some cases.

Private Cloud A private cloud is generally operated inside an organization's network operation. It is generally confined to an organization's internal data center. One can see that since it is limited within the organization, higher level of security is easily achievable. Also, it is easier to maintain and update the systems in this case. As opposed to public cloud, where all the resources are managed by the service provider, in private cloud, these services are assembled together and used by people at the organizational level. Since the organization itself is in charge of the resources, it is managed and updated by the organization itself.

Hybrid Cloud As the name suggests a hybrid cloud is a combination of public and private cloud. In this case, a private cloud is connected to external cloud systems. It provides for a more secured way to share data on public cloud systems. Organizations need public systems for some computation-intensive applications. It thus serves the dual purpose of serving as a private cloud for employees as well as accessing public cloud when need arises.

Community Cloud Community cloud systems are ones where a few or many organizations pool their resources and share a common cloud infrastructure. The cloud system has policies that may be managed by a third party or any of the pooling organizations. They may also be in charge of maintaining and updating the resources and applications.

Having seen the features and architecture of cloud computing, we shall now focus on some of the applications of cloud computing in medical field and pervasive computing.

2.3.3 Applications of Cloud Computing

In Section 2.3.1.2, we discussed about the features of cloud computing, the service models. In this section, we build up on the applications to focus on medical applications using the knowledge we gained from Section 2.3.3.

Cloud computing can provide the virtual infrastructure for utility computing like the Internet-of-things, which integrates monitoring devices, storage devices, analytics tools, visualization platforms, and client delivery. The Internet-of-things is a scenario in which objects are provided with unique identifiers and the ability to transfer data over network. This results in the generation of enormous amounts of data, which have to be stored, processed, and presented in a seamless, efficient, and easily interpretable form where cloud computing comes into picture. To give a better perspective about how cloud is helping the medical field, let's consider an example of conventional method of recording a patient's data in hospitals. The following steps are involved in the process [47].

The nurse records patient's data on a paper in tabular form. This table is then entered by the staff on a computer. The data from the computer is then transferred to a central database terminal, which organizes and makes it accessible. The data is then made available to the doctor and other consultants through the database interface.

At once, we can see the several problems related to this process. It is slow: there is lag between the data collected and data made available to doctors [48]. This system is also prone to human errors; it still depends on human beings to a large extent. Also compatibility becomes an issue, and changes cannot be incorporated because of incompatibility between the handwritten data and the computer spreadsheet.

The way pervasive computing can help in improving this situation is by automating the process. A network of sensors can collect data from the medical devices and then transfer the data collected to the cloud, which organizes, creates alerts and notifications, and makes it ready for the people concerned to view whenever they want to. The entire process now takes much less time and is practically error free since the human element has been removed from the process. The amount of data being collected and the frequency of transmission of the data to cloud can be varied as per the application and necessity.

The sensor nodes have the hardware and software to collect raw data and transmit data through wireless network to the cloud or centralized database servers. Exchange services act like mediators between the cloud and sensors. Its function is to collect data from sensors and route it to the cloud network. It also retrieves the data from the cloud at the request of medical staff. With technological developments, sensors that can be worn by the patients continuously for 24 h monitoring are being put to practice. Such biosensors combined with wireless networks make ubiquitous computing possible and help the health experts get immediate and accurate information on the progress of the treatment.

2.4 CONCLUSION

In this chapter, we presented a brief introduction to nanocomputing and cloud computing. Both of these techniques are very well suited for medical applications, as will be demonstrated in this book.

ACKNOWLEDGMENT

The authors are grateful to Vasudha Goel for her helpful contributions and suggestions during the course of the chapter.

REFERENCES

[1] M.M. Eshaghian-Wilner, Bio-Inspired and Nanoscale Integrated Computing, John Wiley & Sons, Inc., Hoboken, 2009.

[2] Y. Luo, C.P. Collier, J.O. Jeppesen, K.A. Nielsen, E. DeIonno, G. Ho, J. Perkins, H.-R. Tseng, T. Yamamoto, J.F. Stoddart, and J.R. Heath, Two-Dimensional Molecular Electronics Circuits. ChemPhysChem 3, 519, 2002.

[3] C.P. Collier, E.W. Wong, M. Belohradsky, F.M. Raymo, J.F. Stoddart, P.J. Kuekes, R.S. Williams, and J.R. Heath, Electronically Configurable Molecular-Based Logic Gates. Science 285, 391, 1999.

[4] J. Chen, M.A. Reed, A.M. Rawlett, and J.M. Tour, Large On-Off Ratios and Negative Differential Resistance in a Molecular Electronic Device. Science 286, 1550, 1999.

[5] C.P. Collier, G. Mattersteig, E.W. Wong, Y. Luo, K. Beverly, J. Sampaio, F.M. Raymo, J.F. Stoddart, and J.R. Heath, A [2]Catenane-Based Solid State Electronically Reconfigurable Switch. Science 289, 1172–1175, 2000.

[6] H.-R. Tseng, D. Wu, N. Fang, X. Zhang, and J.F. Stoddart, Nanoelectromechanical Switching in a Self-Assembled Monolayer of [2]Rotaxanes on Gold. ChemPhysChem In press, 2003.

[7] M.A. Reed, J. Chen, A.M. Rawlett, D.W. Price, and J.M. Tour, Molecular Random Access Memory Cell. Appl. Phys. Lett. 78, 3735, 2001.

[8] J.K. Mbdindyo, T.E. Mallouk, J.B. Mattzela, I. Kratochvilova, B. Razavi, T.N. Jackson, and T.S. Mayer, Template Synthesis of Metal Nanowires Containing Monolayer Molecular Junctions. J. Am. Chem. Soc. 124, 4020, 2002.

[9] M. Wilson, K. Kannangara, G. Smith, M. Simmons, and B. Raguse, Nanotechnology, Basic Science and Emerging Technologies. Chapman & Hall/CRC, Boca Raton, 2000.

[10] D. Goldhaber-Gordon, M.S. Montermerlo, J.C. Love, G.J. Opiteck, and J.C. Ellenbogen, Overview of Nanoelectronic Devices. Proc. IEEE 85, 34, 1997.

[11] J. Jortner and M. Ratner, Molecular Electronics, Blackwell Science, Oxford, 1997.

[12] Y. Luo, H.-R. Tseng, D.W. Steuerman, J.F. Stoddart, and J.R. Heath, Conservation of Molecular Mechanisms in Solution, Half Devices and Full Devices. Angew. Chem. Int. Ed. (to be published).

[13] J.R. Tucker, Complementary Digital Logic Based on the “Coulomb Blockade”. J. Appl. Phys. 72, 4399, 1992.

[14] J.R. Heath and M.A. Ratner, Molecular Electronics. Phys. Today 56, 43, 2003.

[15] K. Likharev, Single-Electron Devices and Their Applications. Proc. IEEE 87, 606, 1999.

[16] Y. Ono and Y. Takahashi, Single-Electron Pass-Transistor Logic: Operation of Its Elemental Circuit. Electron Devices Meeting, 2000. IEDM ’00. Technical Digest, December 10–13, 2000, San Francisco, CA, 297, 2000.

[17] K. Yano, T. Ishii, T. Sano, T. Mine, F. Muri, T. Hashimoto, T. Kobayashi, T. Kure, and K Seki, Single-Electron Memory for Giga-to-Tera Bit Storage. Proc. IEEE 87, 633, 1999.

[18] D.H. Kim, S.-K. Sung, J.S. Sim, K.R. Kim, J.D. Lee, B.-G. Park, B.H. Choi, S.W. Hwang, and D. Ahn, Single-Electron Transistor Based on a Silicon-on-Insulator Quantum Wire Fabricated by a Side-Wall Patterning Method. Appl. Phys. Lett. 79, 3812, 2001.

[19] K. Yano, T. Ishii, T. Hashimoto, T. Kobayashi, F. Murai, and K. Seki, A Room-Temperature Single-Electron Memory Device Using Fine-Grain Polycrystalline Silicon. International Electron Devices Meeting, 1993. IEDM ’93. Technical Digest, 541, December 5–8, 1993, IEEE, Washington, DC, 1993.

[20] Y. Ono, Y. Takahashi, K. Yamazaki, H. Namatsu, K. Kurihara, and K. Murase, Si Complementary Single-Electron Inverter. International Electron Devices Meeting (IEDM ’99). Technical Digest, 367, December 5–8, 1999, IEEE, Washington, DC, 1999.

[21] M.A. Kastner, The Single Electron Transistor and Artificial Atoms. Annalen der Physik (Leipzig) 9, 11, 2000.

[22] L.J. Geerligs, V.F. Anderegg, P.A.M. Holweg, J.E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M.H. Devoret. Frequency-locked turnstile device for single electrons. Phys. Rev. Lett., 64: pp 2691–2694, 1990.

[23] G. Snider, A. Orlov, I. Amlani, X. Zuo, G.B. Stein, C. Lent, J. Mez, and W. Porod, Quantum-Dot Cellular Automata. J. Appl. Phys. 17(4), 1394–1398, 1999.

[24] C.S. Lent, S.E. Frost, and P.M. Kogge, Reversible Computation with Quantum-Dot Cellular Automata (QCA). Proceedings of the 2nd Conference on Computing Frontiers, Ischia, Italy, May 4–6, 2005.

[25] C.S. Lent, P.D. Togaw, and W. Porod, Bistable Saturation in Coupled Quantum Dots for Quantum Cellular Automata. Appl. Phys. Lett. 62, 714, 1993.

[26] H.K. Raj, A Survey of Cloud Computing. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 4(7), 352, 2014.

[27] M.M. Eshaghian Wilner, Mapping Arbitrary Heterogeneous Task Graphs onto Arbitrary Heterogeneous System Graphs. Int. J. Found. Comput. Sci. 12(5), 599, 2001.

[28] M.M. Eshaghian Wilner, Heterogeneous Computing. Artech House Publishers, Boston.

[29] M. Eshaghian-Wilner, Nano-scale and bio-inspired integrated computing. Oxford: Wiley-Blackwell, 2009.

[30] M.M. Eshaghian-Wilner, A.H. Flood, A. Khitun, J. Fraser Stoddart, and K.L. Wang, "Molecular and Nano scale Computing and Technology," to appear as a book chapter in the edited volume by Albert Zomaya, entitled "Handbook of Innovative Computing," Springer-Verlag (USA), 2006.

[31] D.D. Awschalom, M.E. Flatté and N. Samarth, Spintronics. Sci. Am. 286, 66, 2002.

[32] M. Ercegovac, T. Lang, and J. Moreno, Introduction to Digital Systems. John Wiley & Sons, Inc., New York, 1999.

[33] Y. Ohno, D.K. Young, B. Beschoten, F. Matsukara, H. Ohno, and D.D. Awschalom, Electrical Spin Injection in a Ferromagnetic Semiconductor Heterostructure. Nature 402(6763), 790, (1999).

[34] M.M. Eshaghian-Wilner., A. Khitun, S. Navab, and K.L. Wang, A Nano Scale Reconfigurable Mesh with Spin Waves. The ACM International Conference on Computing Frontiers, Ischia, Italy, May 2–5, 2006.

[35] M.M. Eshaghian-Wilner, A. Khitun, S. Navab, and K.L. Wang, Nano Scale Modules with Spin Wave Inter-Communications for Integrated Circuits. Proceedings of the NSTI Nanotech 2006, Boston, MA, May 7–11, 2006.

[36] X. Zhang and T. Kahveci, A New Approach for Alignment of Multiple Proteins. Pac. Symp. Biocomput., 11, 339–350, 2006.

[37] C. Grasso and C. Lee, Combining Partial Order Alignment and Progressive Multiple Sequence Alignment Increases Alignment Speed and Scalability to Very Large Alignment Problems. Bioinformatics (Oxford, England) 20(10), 1546, 2004.

[38] C. Lee, C. Grasso, and M.F. Sharlow, Multiple Sequence Alignment Using Partial Order Graphs. Bioinformatics (Oxford, England) 18(3), 452, 2002.

[39] Y. Ye and A. Godzik, Multiple Flexible Structure Alignment Using Partial Order Graphs. Bioinformatics (Oxford, England) 21(10), 2362, 2005.

[40] G. Shepherd, Introduction to Synaptic Circuits. In: G. Shepherd (ed.), The Synaptic Organization of the Brain, 5th edition. Oxford University Press, New York, 2004.

[41] B.W. Mel and J. Schiller, On the Fight between Excitation and Inhibition: Location Is Everything. Sci. STKE 2004, pe44, 2004.

[42] A.C. Parker, A.K. Friesz, and A. Pakdaman, Towards a Nanoscale Artificial Cortex. The 2006 International Conference on Computing in Nanotechnology (CNAN'06), CDES, June 26–29, 2006, CSREA Press, Las Vegas, pp. 227–241.

[43] http://wing.comp.nus.edu.sg/pris/AssociativeMemory/HopfieldModel.html (accessed February 23, 2016).

[44] J.J. Hopfield, Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proc. Natl. Acad. Sci. U. S. A. 79(8), 2554, 1982.

[45] M.M. Eshaghian-Wilner, A. Khitun, S. Navab, and K.L. Wang, A Nano Scale Architecture for Constant Time Image Processing. Phys. Status Solidi (a) J. 204, 1931, 2006–2007.

[46] S.H. Bokhari, On the Mapping Problem. IEEE Trans. Comput., c-30(3):207, 1981.

[47] C.O. Rolim, F.L. Koch, A. Sekkaki, and C.B. Westphall, Telemedicine with Grids and Wireless Sensors Networks. Proceedings of eMedisys'08, International Conference on e-Medical Systems, October 29–31, 2008, Sfax Al-Jadida, Tunisia. IEEE Tunisia Section, 2008.

[48] C.O. Rolim, F.L. Koch, C.B. Westphall, J. Werner, A. Fracalossi and G.S. Salvador, A Cloud Computing Solution for Patient's Data Collection in Health Care Institutions. Proceedings of the Second International Conference on eHealth, Telemedicine, and Social Medicine, February 10–16, 2010, St. Maarten, pp. 95–99.

PART II

PERVASIVE WIRELESS COMPUTING IN MEDICINE

3

PERVASIVE COMPUTING IN HOSPITALS

Janet Meiling Wang-Roveda[1,3], Linda Powers[1], and Kui Ren[2]

[1]*Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA*
[2]*Department of Computer Science and Engineering, University of Buffalo, State University of New York, Buffalo, NY, USA*
[3]*University of Southern California, Los Angeles, CA, USA*

3.1 INTRODUCTION

Pervasive computing in hospitals does not come overnight. It started with healthcare providers using computers in hospitals to store patient records. At its early stage, pervasive computing was nothing but a number of desktop computers serving as basic information systems distributed in different locations [1–4]. Different from regular office workers, most clinicians, nurses, doctors, and specialists need to continuously have access to patient information to provide accurate diagnostic results. They have to physically go to various locations to retrieve patient records including magnetic resonance imaging (MRI) images, X-ray images, and other medical documents. The early understanding of pervasive computing research was that this mechanism could eventually intelligently support clinician-centered activities with easy communications among distributed locations and time zones.

Recent years have witnessed a rapid growth of various sensors (e.g., wired or wireless and wearable sensors), mobile platforms that can directly connect to these sensors through Bluetooth, Wi-Fi (e.g., smart phones and smart tablets), and heterogeneous embedded systems that have the ability to record real-time measurement (e.g., smart watch, actigraphy, and fitbit). These ubiquitous computing devices shift the healthcare paradigm from the clinician-centered model to the patient-centered one.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

Today, a number of clinics and hospitals have already demonstrated the advances in early system prototypes including electrocardiography (EKG), photoplethysmogram (PPG), electroencephalography (EEG), and various forms of systems with imaging sensors [5–7]. These equipments and instruments have integrated high-end microprocessors and memories that allow them to become an information hub themselves, not just a measurement equipment. Advances in software (e.g., Android operating systems (OS) for mobile platforms, html, Web-based tools, and application tools) have further furnished the existing hardware to smoothly communicate with each other and with a large group of remote servers referred to as *Cloud*. Cloud computing enables the pervasive computing in hospitals to share centralized data storage, online access to databases, and high-performance data mining and computations to further facilitate accurate diagnosis. Health care is no longer bounded by the walls of hospitals. Telehealth, telemedicine, and home patient monitoring technologies promise to deliver continuous care even outside the hospital walls and increase the overall efficiency and effectiveness of health care.

Nowadays, pervasive computing has a wide range of applications: from wearable sensors and embedded systems to hospital equipments and facilities. It is a combination of ubiquitous computing, cloud computing, and real-time sensing and monitoring. Because of its moderate cost and its ability to collect data in real time, healthcare providers believe that pervasive computing "carry the promise of drastically improving and expanding the quality of care across a wide variety of settings and for different segments of the population" [5]. Because of its wide range of applications, pervasive computing in hospital demands the integration of nursing, medicine, radiology, public health with cloud computing, biomedical instrument design, context-aware computing, human–computer interaction, and artificial intelligence [1].

Pervasive computing revealed two grand challenges related to the data acquired from these systems: (i) Effectiveness and (ii) Privacy and Security. Effectiveness refers to the combination of data reliability and delivery. For example, real healthcare applications require hundreds of thousands of samples per second to create high-quality images or monitor motions and activities in real time. In turn, they demand high throughput and wideband networks to deliver a complete set of measurements in real time. In other cases, the quality of the data collected from sensors in these systems can be compromised due to sensor defects. Considering healthcare givers and staffs rely on these systems for accurate diagnosis, loss in quality due to device malfunction is of great concern. In addition, these sensing systems may enable continuous sampling under situations unexpected by the system designers. The collected measurements may be contaminated by a variety of noise sources. While complete failure of a medical device jeopardizes patient safety, even subtle errors (whether due to noise or systematic errors in the device) can be very problematic if such errors mimic symptoms of a medical condition. Privacy and security is another major concern. Sensing and monitoring systems in health care are used to detect abnormal tissues inside each patient, to record the daily living activities, and to supply data for longitudinal studies. At the same time, these efforts also pose possibilities to violate privacy. With the rise of the adoption of these prototype systems, we expect to see the increasing importance of securing the data. While cloud computing provides an ideal

paradigm in assisting our proposed system design for efficiency and effectiveness, it inevitably introduces new security problems. This is not only because the cloud is an open environment operated by external third parties but also because many measurement datasets are privacy sensitive by nature. For instance, many medical images may include diagnostic results for different patients. Therefore, it is of critical importance to ensure that security be embedded in the envisioned system design from the very beginning so that we can better leverage the cloud computing power, protect confidential data, yet without sacrificing the usability and accessibility of the information.

This chapter will examine the type of pervasive computing techniques in hospital environment. We first introduce sensors and instruments used in the current hospitals. Then we discuss the big data issue caused by real-time sensing and monitoring at the hospital. Cross platform comparison, data compression, and cloud computing are introduced in details to show how these new techniques can be applied to resolve the large data volume issue and data security issue for pervasive computing in hospitals.

3.2 ARCHITECTURE OF PERVASIVE COMPUTING IN HOSPITALS

Pervasive computing has three main components: sensors/instruments, local computers/embedded systems, and data servers. Figure 3.1 provides a detailed systematic view of the pervasive computing environment in hospitals [8]. This is consistent with the new REACTION project being developed at European Union for diabetic patients. For instance, we can divide the pervasive computing environment into two spheres and three groups of servers: patient's sphere, care giver's sphere, the hospital platform, the health information system (HIS) [9], and authentication. Inside patient's sphere, sensors, devices, and instruments are connected to patients through wireless and wired technologies. This patient sphere concept is applicable to operation rooms, intensive care unit (ICU), and regular care unit where wearable sensors including regular medical devices and actuators are seamlessly integrated with local computers and embedded systems for real-time monitoring. In most situations, multiple sensors are connected to patient body to form a body area network (BAN). The BAN provides contextual records for vital physiological parameters and patient's activities. The data acquired from BAN are analog waveforms that are converted to digital signals using analog to digital converters (ADCs). Then the preprocessed digital data are further formatted to be accepted by Open Service Gateway Interface (OSGI) software. This gateway software can manage personalized, patient-centric intervention to automatically adjust BAN devices and instruments. In addition, it bundles OS support on various mobile platforms such as PDA, Apple iPAD, iPhone OS, Android OS, and other supporting software and delivers support directly to point-of-care to achieve patient-centric service. We would like to point out that low-power design, signal processing, system integrations, and multisensor data fusion are still active research topics [8].

Based on diverse functionalities and service types, data servers can be classified as different groups for communication, applications, management, and authentication.

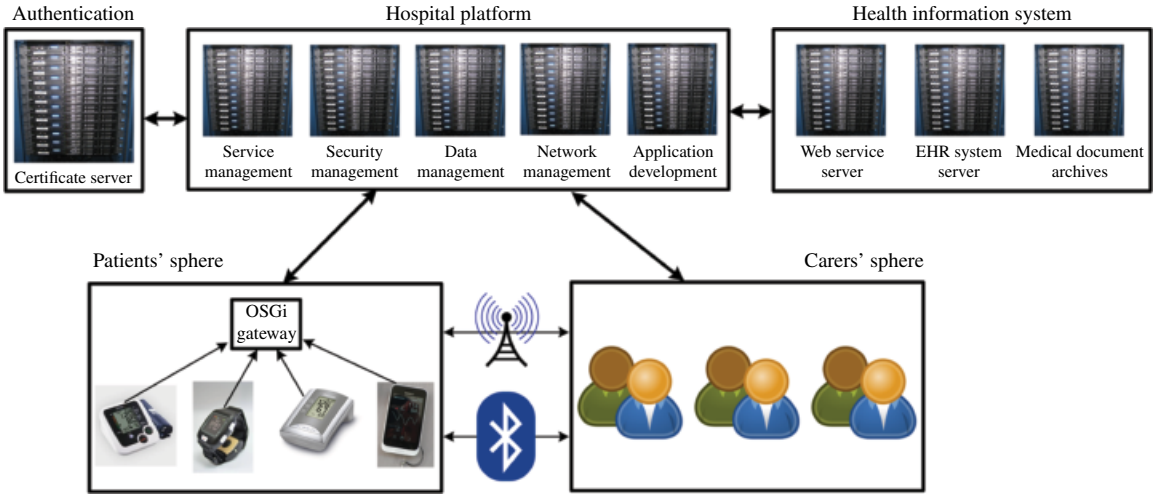


FIGURE 3.1 The systematic view of the pervasive computing architecture in hospitals using new information and computer engineering design technologies [8].

Today, most hospitals allow servers to operate through Web services to achieve scalability. Further, these data servers may be virtualized and can be regarded as part of local or public cloud. We will further discuss the security issues involved in the cloud computing in the later part of this chapter.

Inside hospital platform, servers for applications are dedicated to launch and develop new versions of application software and to execute computing resource demanding tasks. For instance, new versions of image-processing software may be delivered from these servers to local computers at operation room, ICU, or regular patient rooms. Comprehensive tasks, such as image reconstruction, complicated artificial intelligent tasks for data mining, and decision making, require high-performance computing capability and can be deployed to execute on application servers. Figure 3.2 shows an example of ICU and its associated domain model. Each application is defined as one-domain model that is constructed by a series of tasks and rules. Use cardiovascular surgery recovery as one example. Patients usually stay in an ICU for 3–5 days. Continuous measurements of electrocardiogram (ECG), body temperature, blood temperature, heart bump, and actuators that control intravenous injections with antibiotics and pain killers are performed. The feedbacks of these devices and sensors will be managed by application servers following certain domain models. For example, if the blood pressure (BP) is lower than certain threshold, some warning signals will be issued to caregiver or nurse sphere for immediate actions. Additional instructions for medication dose will be provided at the same time to doctors and nurses. Figure 3.2 (right) indicates such a task flow following rules in the domain of cardiovascular surgery recovery. The domain model is defined by users (doctors and nurses), rule sets, device measurements, and patient status. Model-based design and tools [10–17] can be employed to build and construct these applications.

Servers for data management offer storage and transferring from one hospital to another and provide high-level model-based architecture for applications, data fusion, and other data-related services. These servers orchestrate the sequence of actions among different embedded systems or local computers for devices and sensors. Servers for communications include network management servers and security servers. The network management server has interfaces for external Web service and exchange data among different servers through physical communication channels. The security management server maps the user and client to trustable domains or servers.

The carers' sphere receives personalized patient data from data management servers and communication and security servers. The data are generated from devices, sensors, instruments, actuators, embedded systems, and local computers at the patient's sphere. The carers' sphere may not be physically located in the same building. It is a virtualized concept and may include healthcare professionals working in another hospital ward. Such a concept allows medical data sharing among different outpatient clinics, specialists, and professional caregivers to coordinate in medication adjustment and rehabilitation program generation. This also enables efficient risk assessment and health status monitoring at emergency response centers. Contextualized health information may also be shared with patients and their family members [8].

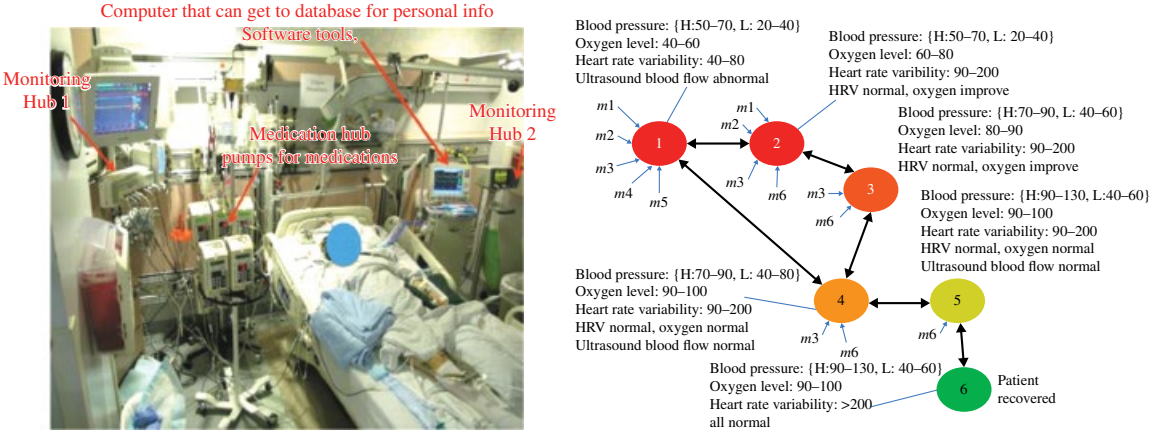


FIGURE 3.2 Intensive care unit for cardiovascular surgery recovery (left) and task flow for the domain model (right).

Health information systems (HIS) [9] refer to servers that store and manage health-related records and activities for patients. This also includes any decisions, activities, and diagnostic-related information from doctors, specialists, nurses, caregivers, and organizations at the health sector. HIS [9] can perform several different types of functionalities including disease surveillance, laboratory information collection, organization-level routine processing, human resource management, and hospital patient administration. A successful HIS [9] system can provide health statistics with age, gender, and other parameters to influence policy making (e.g., insurance), health- and care-related research, and resource match and allocations. Due to the large volume of data collected at HIS [9] systems, it is important to find out how to effectively collect data, how to guarantee the quality of data, and how to efficiently search the right information.

Servers for authentication provide network service to applications and individuals to authenticate the access or use of information owned by the pervasive computing environment at hospitals. This includes account names and passwords for all users (e.g., patients, doctors, specialists, nurses, staffs, and anyone related to hospitals). The users usually send a set of valid credentials to receive a cryptographic key that can access to various information resources [18]. Authentication is thus referring to this determination process whether a user should be granted with a privilege or not. This is an important measure to keep private information safe and to prevent nonauthorized users from getting information or access to the hospital resources. Three popular protocols related to the authentication of network service include terminal access controller access-control system (TACACS), Remote Authentication Dial-In User Service (RADIUS), and multifactor authentication (MFA). TACACS employs a centralized server to provide a family of protocols for remote access control through networks. Developed in 1984, TACACS was commonly used in UNIX networks. RADIUS, first introduced in 1991, is a network protocol with the Internet Engineering Task Force (IETF) standards [19] that authorize, authenticate, and manage user accounts. As a background process on a UNIX or Microsoft Windows server, RADIUS is currently a mostly used client/server protocol that can approach applications and all gateways for network access and services. Therefore, it is regarded as a ubiquitous protocol that is applied to networks with digital subscriber line (DSL), virtual private network (VPN), and Web servers [20–22]. By submitting knowledge factor (e.g., passwords), or possession factors (e.g., ATM cards), or inherence factors (e.g., biometrics), a user can get access to devices, sensors, instruments, mobile platforms, embedded systems, and local computers through MFA method. It is clear that MFA is the key method for access control of devices, sensors, and other components at patient's sphere. RADIUS is the main protocol for the rest of the hospital-computing environment. We will further discuss data volume issues and security challenges in the later part of this chapter.

3.3 SENSORS, DEVICES, INSTRUMENTS, AND EMBEDDED SYSTEMS

According to the World Health Organization (WHO), medical sensors, devices, instruments, and their associated embedded systems or computers or mobile controller are all referred to as "medical device" [23]. In addition, "medical device" also means any

"implants, software, apparatus, machine, aids for persons with disabilities, reagent for in vitro use" [23]. The purpose of medical devices is to support and sustain human life through diagnosis, prevention, monitoring, treatment, or alleviation of disease. According to WHO's estimate, more than 10,000 types of medical devices exist around the world. However, not every type is available to patients. Economic differences among nations and quality of health care and amount of investment in health facilities are the key factors that influence the accessibility of these medical devices. In this section, we review wearable devices and imaging devices as the two main types of medical devices that play important roles in pervasive computing at hospitals.

3.3.1 Wearable Sensors

Wearable sensors are becoming the mainstream type of medical devices in recent years because of the possibility of using these sensors at both hospital and home. The development of wearable sensors will further lower healthcare cost, allow patient care to be location independent, and ultimately enable patient-centric care outside hospitals and clinics. In this book, we focus on wearable sensors at hospitals. Two types of wearable sensors are used in the hospital and clinical environment: vital sign sensors and activity monitoring sensors. For example, vital sign sensors include ECG, heart rate (HR), respiratory rate (RR), oxygen saturation (SpO_2), photoplethysmogram (PPG), blood glucose (BG), and BP. Activity monitoring sensors are mainly based on accelerometers and are used often for activity recognitions and rehabilitations. Some of the wearable sensors integrate several different sensors into one device. Multiple sensors may be used to measure physiological development or to support complicated diagnosis and decision-making process. We will discuss one case study: wearable sensor-based vagus nerve monitoring.

3.3.2 Case Study: Wearable Sensor-Based Vagus Nerve Monitoring

Wearable technology has been widely used in clinical context such as disorder detection, treatment efficiency assessment, home rehabilitation, and other healthcare research. One particular focus area is stress management since many people have experienced stress and have been affected negatively during their surgery recovery. It has been found that activating the vagus nerve promotes a relaxation response to mental stress. It is possible to monitor vagus nerve activity using simple, inexpensive, and readily available technology such as wearable technologies, mobile platforms, and real-time data-driven tracking and management. We present this case study to show how to use wearable technology and mobile platform to monitor vagus nerve activity and increase it when needed or as desired in order to promote relaxation and better health. Enabling vagus nerve activity has been found to be very useful in wound healing, postcardiovascular surgery recovery, and post-traumatic stress disorder.

This case study selected Zephyr BioPatch as a wearable sensor unit [24]. BioPatch collects patient HR, RR, ECG, and positional and activity information in real time. It transmits patient vital signals to monitoring state through encrypted radio signals

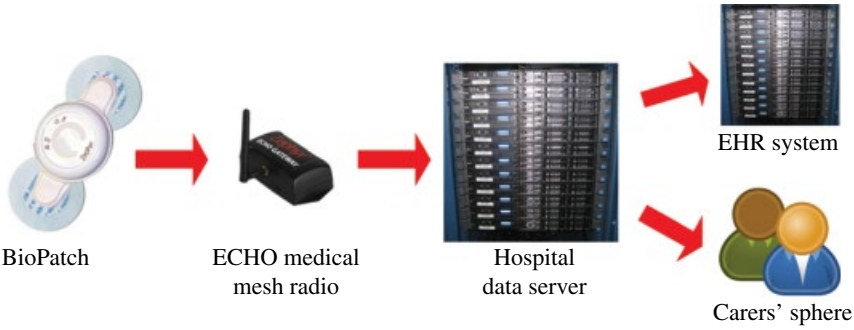


FIGURE 3.3 Wearable integrated sensor and system at hospitals [24].

with mesh medical radio. Nurses at caregiver sphere receive visualized waveforms and signals every minute. Figure 3.3 displays the BioPatch-integrated sensor unit with mesh medical radio for hospital application. At the front end, wearable sensors such as BioPatch-integrated sensors provide measurements of patients' RR, HR, body movements, and other basic information. These wearable sensors are low power and can work over 24 h without charging to create baseline measurements for patients. An embedded system in patient room such as a mesh radio can establish reliable communications between the BioPatch sensor and a secure application server, which is interconnected to the central monitoring station at caregiver's sphere and electrical health record (EHR)/personnel health record (PHR) at HIS [9] discussed in Section 3.2. Applications for stress management can be directly implemented on application servers. To illustrate, application servers accept measurements through wireless mesh radio communications with wearable sensors. Equipped with multiple cores and wired data links, application servers communicate with local data centers and enable basic data analysis capability locally. HIS [9] servers offer data storage for patients and real-time data monitoring and analysis for healthcare providers. Text messages and warnings can be directly displayed on health giver's computers and smartphones through an SaaS. Figure 3.4 shows the waveform signals from BioPatch on computer screens at nurse stations in caregiver sphere.

The vagus nerve-monitoring application integrates stress check with relaxation exercise. It performs real-time stress check while the user is following the recommended relaxation program. We emphasize stress intervention through relaxation program (i.e., breathing training programs) to form a feedback loop, which can help users to lower down their stress to a "safe" level [26]. It prevents the users interrupting their relaxation programs without realizing that their stress levels have not reached to an optimal level. A local computer or a mobile platform such as patient's smart phone can display this relaxation exercise program to form an effective, real-time stress management. Through mesh radio communication, BioPatch data are transmitted to data servers to facilitate off-line data analysis to further monitor users or patient stress level. At application servers, we can launch calibration algorithms based on symbolic data representation to process breathing waveform data in real time and reduce the data storage. These types of algorithms employ a pruning algorithm called

FIGURE 3.4 ECG display (left) and warning level (right) at nurse station in caregiver's sphere [25].

distance filter with starting symbol requirement (DFSSR) to provide fast feedback to real-time relaxation application. This feedback is critical to know how well the subject is adhering to the prescribed breathing pattern for relaxation exercise. A 2D visualization is also included to display stress management progress. This approach can also be launched onto pervasive computing platforms to provide data sharing and analysis among caregivers and patients.

Several key technologies are adopted to implement such an application. For example, we use Google's Android OS to create application programming interfaces (API) between wearable sensors and smartphone. On application servers, various data analysis and mining approaches are exploited to provide reliable stress-level estimation. For example, we provide a set of HR variability evaluations to justify ECG noise. We also use QRS detection algorithms [27, 28] to analyze HR variability. QRS is a name for the combination of the three graphic deflections on an ECG waveform. App data analysis and feature filtering are installed on smartphone platform. Index-equipped database storage on local data center to facilitate data analysis and monitoring. The smart phone stress management app has been designed as a local data analysis computing resource, an interface for the user to access complex functions such as configuration setting. It is also the interface with the cloud. It collects data (i.e., ECG and RR) from wearable sensor and displays some key parameters on the phone at real time. In this case study, we apply a yoga breathing exercise for this stress management app to help patients and users to relax and reduce their stress level.

The stress management app has been implemented on Samsung Galaxy III with Android 4.3 OS. When the user is doing the exercise, the user's breathing waveform is also being monitored. The DFSSR achieves reduction rate of 91.41, 87.11, 98.27, and 43.75% for each stage and 97.06% totally. Therefore, the size of hash tables used to map symbols to cluster gets reduced with a ratio of 97.06%, which is feasible for a fast implementation. Figure 3.5 is the visualized representation. Horizontal axis is time at 10 s, which is the length of one breathing cycle. Vertical axis is breathing waveform from stage 1 to 4. Figure 3.4 visually describes the changes of testing subject's stress level. Testing the subject is done under high-level stress during the label "high stress level" period on the left side. We could observe that red color occupies a large portion on the left side. Then, stress level starts to decrease around the middle along the horizontal axis during the label "reduced stress" period where

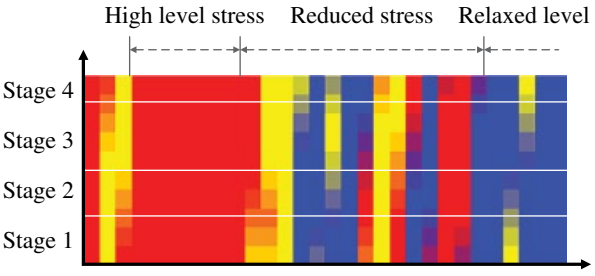


FIGURE 3.5 2D color correlation visualization.

the color starts changing between yellow and blue. Finally, the stress level goes back to a healthy state at last during the label "relaxed level." The stress level reduction process demonstrates that the stress management program can successfully help the user relieve stress via a breathing exercise.

This case study shows that any research or projects in the pervasive computing area will exceed beyond research in biomedical monitoring and data-related computer information. Pervasive computing in hospital, as a research topic, also enables multidisciplinary educations at hospitals, clinics, universities, industries, and local communities. It also facilitates the potential development of smart homes with the capability of health monitoring. However, one of the main challenges is the sensor accuracy and data calibrations. As one of the first ones in the data-driven stress management research area, it is fairly important to create benchmarks for future biomedical research.

3.3.3 Pervasive Computing in Medical Images

While much is known about magnetic resonance imaging (MRI), X-ray, ultrasound, and other cancer detection techniques, little has been done to automatically correlate data collected by these different sensing systems to reach an accurate diagnosis. Erroneous sensors, inaccurate analog-mixed signals, noisy processing, and faulty circuit and system designs can all contribute to misinterpretation. Consider breast cancer as an example. While one in eight women has breast cancer in North American population, 35% of digital X-ray screenings for early detection of cancer are inaccurate. On the other hand, many other methods are available, some of which depend on metabolism and have shown promising results in early stage detection.

Ultrasound: As is well accepted, ultrasound imaging is becoming an important technology and has been widely used in medical diagnostics for regions invisible to human eyes. Compared with X-ray and an MRI, ultrasound offers several advantages such as inexpensive, noninvasive, and nonradiation exposure [29]. Furthermore, the ability of real-time imaging and display makes interventional procedures more effective by providing instantaneous visual guidance [29]. However, low resolution and noise are still the two most urgent drawbacks waiting to be solved for modern ultrasound imaging systems. In this section, we focus on ultrasound frequency and device bandwidth of modern ultrasound imaging systems by addressing the basic physics of acoustic signal propagation and working principles of ultrasound imaging system hardware. Through this discussion, we explore the limitation of modern ultrasound imaging hardware in processing the received ultrasonic signal. Figure 3.6 shows a block diagram of hardware components from Texas Instruments Inc. As displayed in Figure 3.7, the ultrasound imaging system is a combination of transducer, circuit, and display.

For a complete scan, the signal starts from the beamformer control unit [30, 32], which generates a pulse in digital form. Then, the digital pulse goes through the standard signal chain including digital to analog converter (DAC), low pass filter (LPF), and linear amplifier becoming analog pulse in the form of a voltage. Next, the analog pulse signal from standard signal chain will be transformed into acoustic pulse signal by a transducer and then transmitted into the human body. After the acoustic pulse has been transmitted, the transducer will be turned to receive mode to

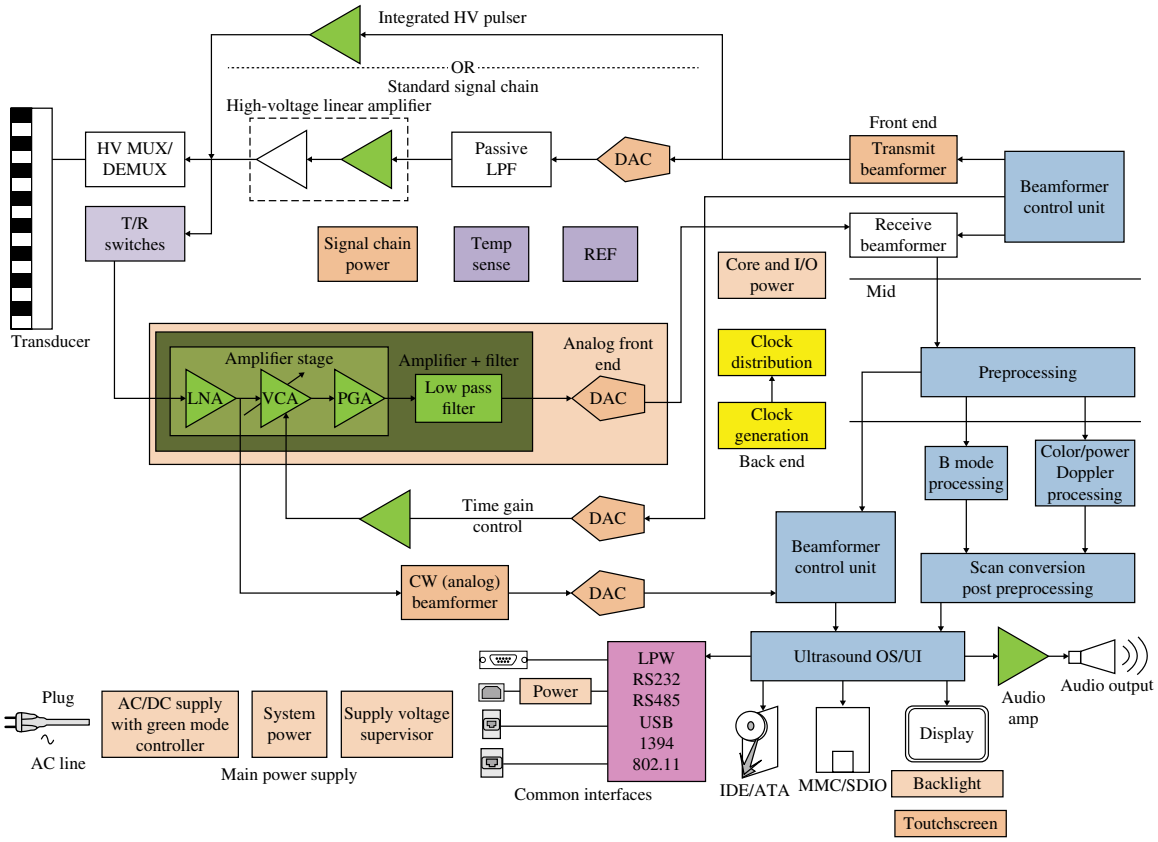


FIGURE 3.6 Block diagram of ultrasound imaging system by Texas Instruments Inc. [30].

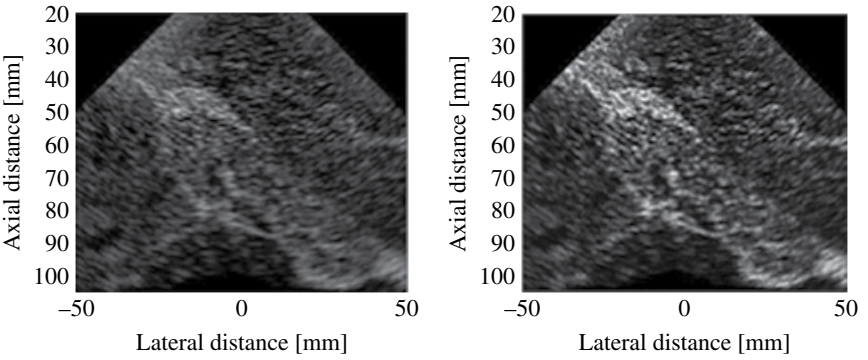


FIGURE 3.7 Preliminary results at the output analog front end (before digital display) from the ultrasound platform implementation on Matlab to demonstrate different noise injects caused differences in displays [31].

immediately receive any acoustic signal (echo) and transform it back to an electrical signal in the form of a voltage. The transducer works both as transmitter and receiver by transforming the acoustic signal to an electrical signal and vice versa. Then, the transformed analog echo will be amplified and converted into a digital signal by analog front end and processed to form the ultrasound image. For example, the transducer usually consists of an array of piezoelectric transducer elements and operates at a main frequency range between 2 and 15 Hz [32]. The overall functionality of the analog front end is to convert the analog signal to a digital one. It includes a low-noise amplifier (LNA), time gain compensation (TGC), passive low-pass filter, and an ADC. Since received echoes are very weak due to acoustic attenuation, the analog signal is first amplified by the LNA to a normal scale. As echoes received at different time are reflected at different depth, attenuation of those echoes is also different. Thus, echoes received later should be amplified more than those received earlier. The TGC works to normalize echoes received at different time by using variable gain changes according to time. Next, the low-pass filter filters out high-frequency signals that are usually noise. Finally, the analog signal is digitalized by the ADC.

The digital part of the ultrasound method, including beamformer [30] and preprocessing unit, is discussed later. The ultrasound image consists of multiple scanning lines to form the image. For each scanning line, multiple transducers are used. Thus, echoes received by each transducer go through a different length inside the tissue. As a result, signals from those transducers need different delays to synchronize the echo phase. The beamformer works as a synchronizer, and the synchronization strategies are predefined according to the scanning mode used. Then, echoes from different transducers are summed up after synchronization. The preprocessing unit is usually implemented using digital signal processor (DSP) or field-programmable gate array (FPGA) to form a different image mode. For example, in B-mode ultrasound imaging, all scanning lines from the beamformer are combined with an appropriate vertical interpolation. The amplitude of signal is also compressed in order to fit into the dynamic display range [32]. Finally, all signals are ready to display on a screen.

Some possible causes may introduce noise and reduce resolution from the hardware perspective. First, the transducer stress frequency has a significant impact on the signal amplitude when Doppler effect occurs. Since speckle noise is generated during acoustic propagation inside tissue, frequency response at nonresonance frequency suffers from both speckle noise and transducer. Once the signal has been amplified by the LNA, some noise may become a dark or bright spot, misleading the diagnostic decision. This is different from the existing modern ultrasound imaging modeling research that only models the acoustic field [33].

A mammography-embedded platform for X-ray processing: We use complementary metal-oxide semiconductor (CMOS)-based sensor array for full-field digital mammography (FFDM) as an example to discuss sensor array design and other components to examine sample acquisition and data processing through this system. The performance metric includes bandwidth, signal-to-noise ratio, and information reliability throughout this system. Figure 3.8 shows a block diagram of a mammography system including X-ray source, platform to hold sample, sensor array, amplifiers (integrated with sensor board), buffers, and ADC converter. Currently, the conventional 2D mammography is the most popular approach to detect early-stage cancer, that is, breast cancer. However, it is well known that the 2D mammography has limitations. In cases of breast cancer detection, a recent report shows that it misses 10–30% of breast tumors [34] due to anatomical noises caused by the overlaps of breast tissues under 2D projections. One effective way to avoid anatomical noises is by using higher-dimensional approaches, that is, 3D and 4D computed tomography (CT) [35] and tomosynthesis. Still, the major challenge is the dramatic increase in the data volume: for 3D tomosynthesis, 15 frames of images are required compared with 2 frames for the traditional 2D ones. This is equivalent to 7 × increase in the data volume, a significant increase for imaging using cone beam CT (e.g., an excessive 300 projections) [36]. High data volume delays data process and storage

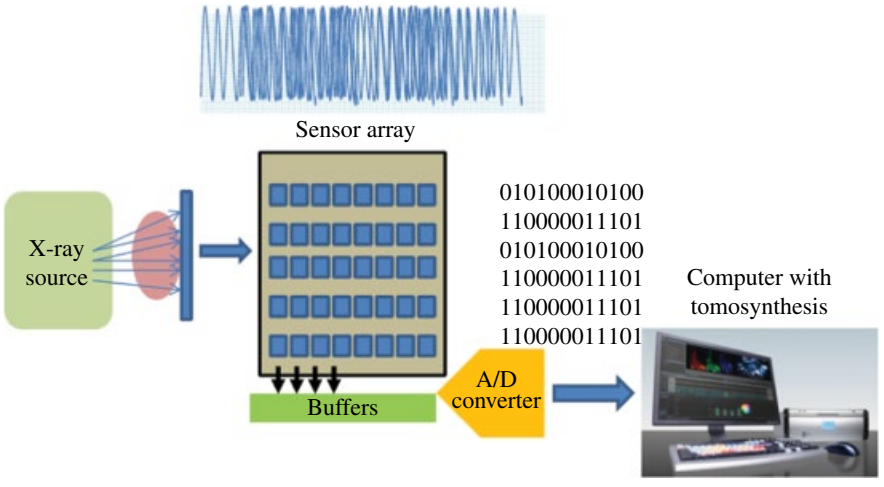


FIGURE 3.8 Implemented mammographic system.

speed and thus degrades the performance of high-dimensional CT. In addition, there is an increase in the X-ray dose as well. For the "average" sample such as a breast (compressed thickness of 5.0 cm and a 50% glandular fraction), a digital breast tomosynthesis (DBT) acquisition results in an 8% increase in X-ray dose comparing with 2D ones. For high-density samples, the increased amount can be as high as 83% [37]. Therefore, how to effectively reduce the data volume without escalating X-ray dose is the key to the new cancer detection system design.

Two potential approaches include how to use multiple low-radiation dose low-quality images to reconstruct a good quality one and how to use existing ultrasound scans to predict mammography images. The first approach will shed lights on how to reuse image data information collected under different conditions from the same system to create a correct image. The second approach will help us to recover missing information through different medical measurements to accommodate fundamental flaws of mammography systems. For example, we can achieve low dose design using high dynamic range (HDR) imaging in the image reconstruction. As an emerging field to introduce high impact in medical imaging research, HDR is based on the fact that light intensity increases linearly throughout the integration time after the initial reset signal. Refer to the "time-pixel input light intensity" graph in Figure 3.9. In every cycle, each integrating-type image sensor receives constant light intensity.

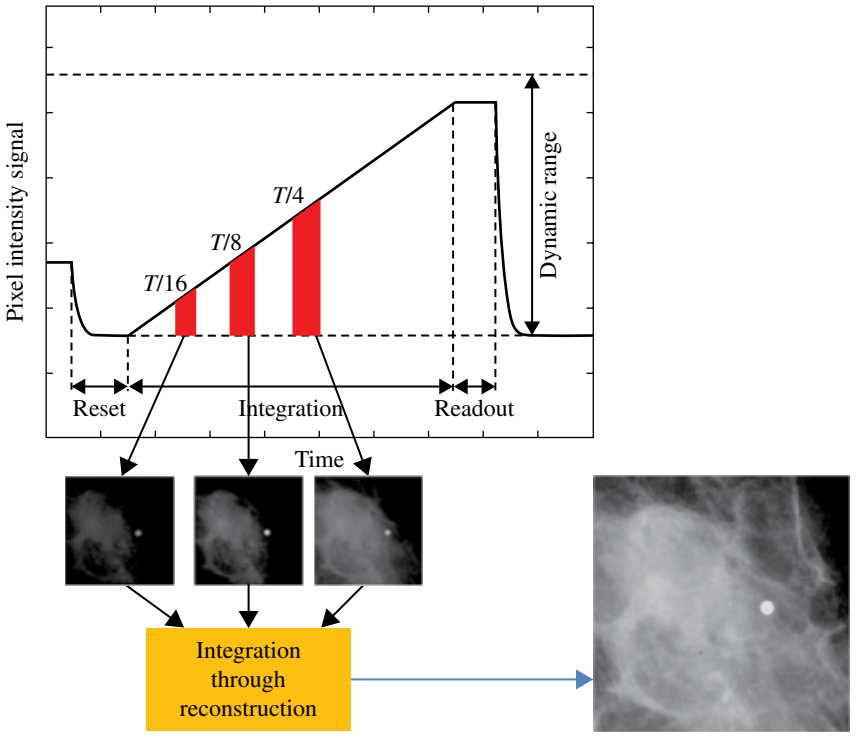


FIGURE 3.9 Low-dose design with image integrations.

The ratio between the illumination that saturates the pixel and the minimum detectable illumination is defined as the dynamic range (DR) of the image sensor. However, noise introduced in the pixel readout circuitry may overwhelm the illumination in low-light intensity situations. Therefore, expanding DR is an important measure to obtain high-quality images. Existing approaches for HDR include reducing the noise floor to capture darker scenes, increasing the saturation point, and enabling the capture of brighter scenes. Several design techniques can further improve the quality of HDR. For example, low noise circuitry, carefully engineered photodiode reverse saturation current, and other improvements can help lower the noise floor. We can also increase the upper bound of the DR by various image-acquiring schemes including logarithmic, a combination of linear and logarithmic, well capacity adjusting, dual or multiple sampling, and multiple integration time [38–40]. After the image alignment, the preliminary HDR image and the maximum a posteriori produced data file are inputs to the reconstruction procedure, which will recover the image based on the three low-exposure ones.

An effective way to compare images from MRI, X-ray, ultrasound, and other cancer detection techniques is to create a cross-examination platform that can perform data comparison throughout various measurement instrument platforms. For example, real test images from the samples will be taken from the mammography system and ultrasound. Reversely using the implemented platform, one can deduct ultrasound image results from mammography images and vice versa. A cross-examination platform can take digitized signals from multiple measurement systems such as X-ray, MRI, and ultrasound images. The data from each will be normalized and compared to provide correlations between measurement methods with the occurrence and stage of the cancer and other factors. This comparison provides the basis to predict the results of other measurement systems from the measurements of one in order to avoid errors and misinterpretations. Instead of using images at the output of traditional imaging systems, the cross-examination platform extracts reliable signals in the existing systems determined by the comparison of multiple measurement systems and uses these to predict by simulations the cancer cell growth rate, location, and size. During the data comparison, we automatically and adaptively adjust the measurement systems to improve data accuracy. Built upon medical image data correlations for different modalities, we may improve the cancer detection rate by mining medical image data, by integrating information from different resources or various medical measurements, and by offering solutions to large-volume medical data management and storage.

3.4 DATA ACQUISITION IN PERVASIVE COMPUTING

In this section, we will discuss the data acquisition design in pervasive computing. For every generation of high-end CPU, the power consumption increases by 2×. While we scale down transistor feature size, our processors are also becoming less power/energy efficient. So where does the energy go? Bill Dally in his recent talk [41] pointed out that "data movement," "driving long wires," "accessing arrays," and "control overhead" are the four major factors that caused energy waste in CPUs. One solution he

suggested was stream computing. IBM's cell processor is one such solution. Viewed as a major breakthrough in image processing for game console development in 2005, the cell processor demonstrated 197 GB/s data flow at 3.2 GHz for a synergistic processing unit (SPU) [42, 43] with 70 W average power consumption. This allows high-definition display for medical images in real time. Thanks to a rapid growth of wired or wireless sensors for various applications, CPU and processors have been applied as part of the data acquisition systems in a number of areas including communication, military tracking and monitoring, health care, physical/biological/chemical experiments, environmental testing, and transportation carriers [42].

Figure 3.10 shows a diagram for a typical data acquisition system. It has a sensor with analog mixed signal front end and stream processor. The performance of these two components is very different. Most analog front ends consume 2/3 of the total chip area. While the power consumption is in the microwatt to megawatt range, the ability to sample and process data is a lot slower than digital processors. For example, a typical 24-bit ADC at Texas Instruments Inc. [44] is capable of 125 ksps (sample per second), which leads to 3 Mbps data processing speed. With 197 GB/s of cell processor SPU, the analog front end is several orders of magnitude slower. This means that with the current stream processor capability, we can consider real-time control of the analog front end to obtain useful samples.

The term "useful samples" refers to the most important information embedded in the samples. It is well known that most of today's data acquisition systems discard a lot of data right before it is transmitted. For example, we use JPEG to reduce data amount right before transmission to avoid lengthy communication times. This section presents a cutting-edge architecture that is data/information oriented and allows us to reduce the amount of data from the very beginning. Figure 3.11 demonstrates a proposed architecture for the new data acquisition system. Note that the data are reduced at the analog front end, thus the stream processor receives only a fraction of the total amount of original data. With a reduced amount of data, less energy will be consumed in "data movement," "driving long wires," "accessing arrays," and "control overhead" in the stream processors.

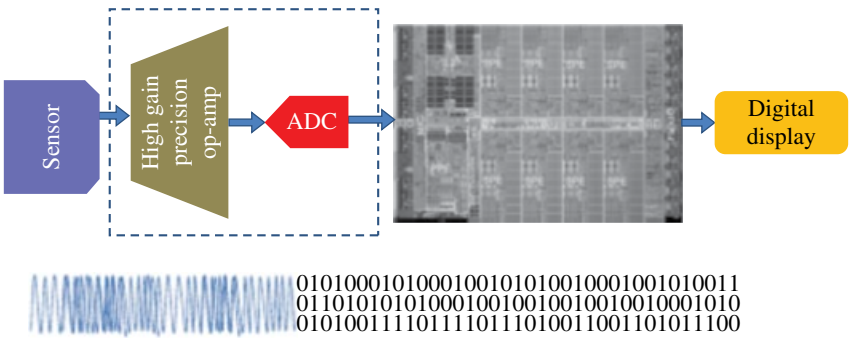


FIGURE 3.10 General block diagram for data acquisition system (the die photo is from cell processor [43]).

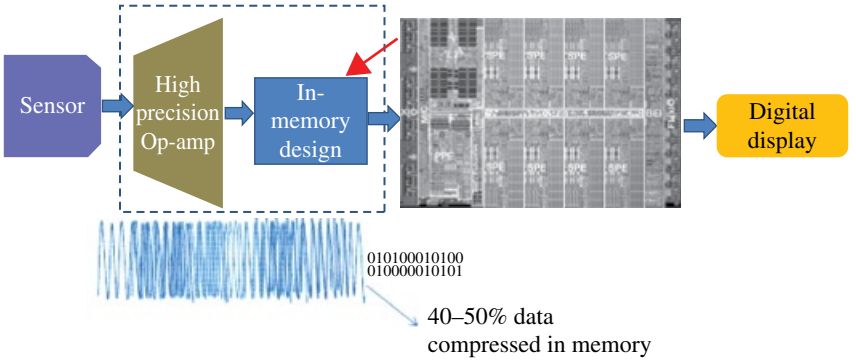


FIGURE 3.11 A new architecture with an analog mixed single-interface design that can reduce the amount of data. The die photo is from IBM cell processor [43].

The advances in nanometer electronic systems, compressive sensing (CS)-based information processing [45–55], and stream computing technologies have empowered us with great potential in creating novel hardware/software platforms having fast medical data processing capability. Driven by this new technology development, the goal of this section is to discuss a high speed adaptive design for information (ADI) system that leverages the advantages of feature-based data compression, low-power nanometer CMOS technology, and stream computing. In particular, we address practical considerations that arise in implementation for the mixed signal interface circuits and how this will integrate with stream processors to achieve an adaptive and low-power data acquisition architecture. We introduce a level-crossing sampling approach to replace Nyquist sampling. We discuss a cutting-edge front end circuit that combines the level-crossing concept with random selection matrices. The circuit performs digitization only when there is enough variation in the input and when the random selection matrix chooses this input. We also discuss mixer functions and random selection functions to understand their capabilities in reducing sensing system random errors and systematic errors. One can implement the discussed sampling concept as an in-memory design to replace the existing ADC and data compression mechanisms.

3.4.1 Overview of Analog-to-Information Converter Concept

Let us begin with a brief review of CS theory [45–55]. An example image can be presented as an $N \times M$ matrix. By using a random selection matrix, we can generate a new matrix:

$$y_{m\times M} = [\Phi]_{m\times N}[X]_{N\times M} \tag{3.1}$$

Because $m \ll N$, we reduce the total data amount. Different from JPEG and other nonlinear compression algorithms, CS linearly reduces data and preserves key features without much distortion. This is the key reason why CS can be applied to the

front end of a data acquisition system instead of right before data transmission. One such example is a "single-pixel camera" [56]. The camera performs random selection on the sampled object. Thus, less data amount will be generated by the camera and subsequently enter the following data acquisition system. Depending on the sparsity of sampled data, the average data reduction of CS is about 50%. If we use joule per bit as an energy estimation, this indicates that this compression algorithm may lead to a total energy/power reduction for the following data processing architecture.

While CS provides great potential, it requires careful implementation. For example, it is still being debated whether data created by a single-pixel camera can provide good randomly sampled data. Over the past several years, much effort has been expended on analog-to-information conversion (AIC) [57–62], that is, to acquire raw data at a low rate while accurately reconstructing the compressed signals. The key components under investigations were ADCs, random filtering, and demodulation. Hasler et al. [63–67] were the first to apply CS to pixel array data acquisition systems. Instead of directly extracting pixel array data to column buffers followed by processing with ADC converters, they used random selection measurement matrices to regroup the pixel data, which are then fed into an ADC converter at a much lower sample rate. They implemented the random selection measurements and multipliers using analog components. Such designs choose a "heavy" analog front end to reduce the sample rate and data amount at the ADC converter. In the traditional data flow, the ADC converter is placed right after the pixel array. That is, the pixel data are directly digitized at the Nyquist sample rate. When a CS algorithm is applied, the ADC converter is placed after the random selection/demodulation and the sample rate is significantly slower. Even though CS algorithms help reduce the sample rate of the ADC converter, it comes with a price. It requires an analog front end to achieve randomized measurements which, in turn, leads to large analog computing units at the front end. These components are cumbersome and slow. For example, an analog multiplier works at 10 MHz with over 200 ns setup time. While most elements in the front end use 0.25 μm technology node, some exploit 0.5 μm technology node (i.e., floating gate technology to store random selection coefficients). By using CS to reduce the sample rate of the ADC converter, it appears that we are moving away from the current technology trend (i.e., smaller feature size transistors to achieve higher speed and lower power). Instead, we rely heavily on analog designs and computations, which have difficulties in scaling. Little is known on how to build circuits that can create "good" measurement matrices. Here, "good" not only refers to effective selection matrices but also includes circuit implementation costs such as power and space requirements. In addition, the high complexity of reconstruction algorithms demands high-performance computing capabilities.

3.4.2 Level-Crossing-Based Comparator Array Design

Medical images have different pixel value ranges. For example, pixel values in a breast image may range from 0 to 4,000 compared with lung images from 0 to 25,000. After examining many images, it may be that the lung images require 15-bit quantization while the breast images require 12-bit quantization. This 3-bit difference

leads to 87.5% reduction in the number of comparators (15-bit requires 215 comparators while 12-bit requires 212 comparators), which in turn greatly reduces the power consumption by the comparator array. Additional studies on a large number of images reveal that dense tissues lead to lower pixel values. Higher pixel values contain information about calcification spots (e.g., white spots in the mammography images), so we should focus on the accuracy of the higher voltage level of the comparators. It is possible to reuse the lower-voltage level comparators to refine the lower voltage level, which improves the resolution. To summarize, the application decides power consumption, hardware cost, and image quality.

Level-crossing-based random selection is introduced to quantize the prior knowledge of voltage level and mixing functions or random selection matrices. This new scheme is different from the regular Nyquist sampling theorem and the level-crossing ones. To illustrate, let us first discuss the difference between the Nyquist scheme and the level-crossing one. Refer to Figure 3.12a–c. Figure 3.12a is the regular Nyquist sampling scheme where a clock with period Tclk controls when sampling will occur [57]. Figure 3.12b displays the level-crossing sampling scheme [57]. Here, a sample is accepted only if the input signal crosses one of the predefined voltage levels equally spaced with ΔV. Unlike Nyquist sampling, the time passed between two samples (refer to A and B points and ΔT in Fig. 3.12b) depends on the signal variations instead of the clock period. Figure 3.12c demonstrates the proposed level-crossing-based random selection. In addition to the level crossing, we only perform sampling when the random selection matrix would randomly select this sample. For example, if $V2$ is significantly different from $V1$, the level-crossing scheme shows that this is a potential new sampling point. However, if the random selection matrix has zero value with regard to this particular sample point, we would bypass it. Therefore, a sample is taken only if it is different from the previous sample point and it is randomly selected.

3.5 SOFTWARE SUPPORT FOR CONTEXT-AWARE AND ACTIVITY SHARING SERVICES

In Section 3.4, we discussed wearable devices and medical imaging instruments that help monitor vital signs, physiologic trends, and other important parameters. Today, most physicians are highly mobile: they may visit several buildings within a very short period of time. At the same time, it is important for them to continuously monitor patient status to decide when to delivery care and at what level. On the other hand, real time, continuous data transmission to caregivers' sphere may also generate large volume of alerting signals, which may comprise caregiver's focus. One strategy is to use ambient displays for selected important items such as aesthetical recovery, patient urine amount, IV inject dose, etc. Of course, it is rather challenging sometimes to provide different ambient displays for a complicated situation where careful and precise choices are needed regarding what information to be displayed and when [1, 68, 69]. In a number of situations, it is fairly important for physicians to be aware of where they are, who they are with, and what time frame they have.

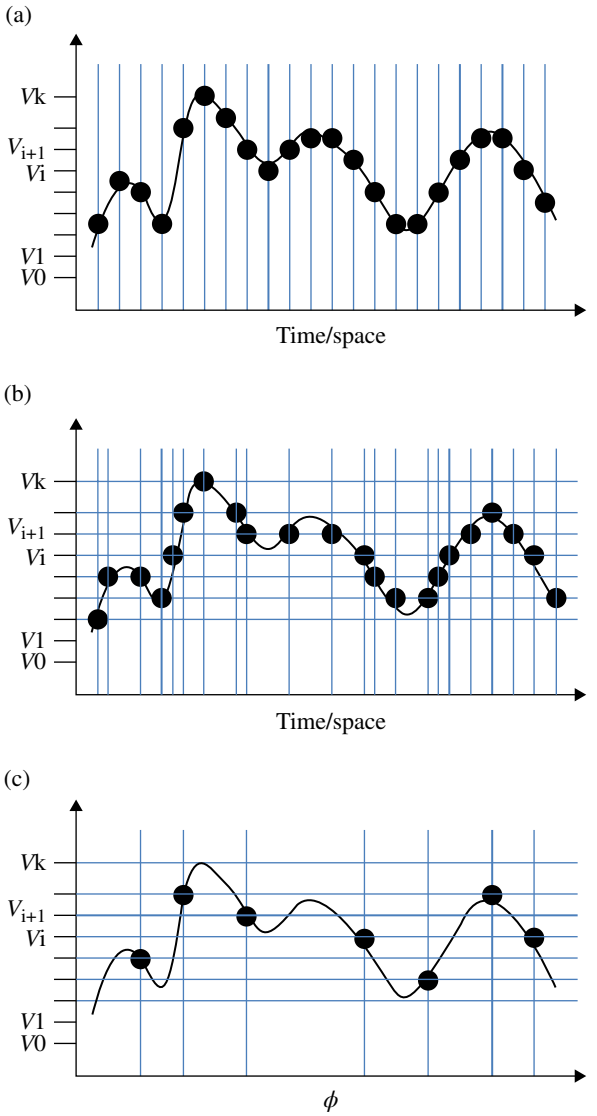


FIGURE 3.12 Different data sampling schemes [57]. (a) Nyquist sampling, (b) level-crossing sampling, and (c) level crossing-based random sampling.

Contextual information such as location, time, and resources sometimes decides the actions a physician may take in the next moment. For example, when a physician is visiting a patient at an ICU, at that moment the patient's medical record on the computer in that ICU room will affect the physician's decisions on patient medication for the next couple of hours [1, 70]. Studies have also shown that caregivers around one clinician or physician can also influence the physician's decisions.

For instance, one nurse reported some detailed observations of one patient to the visiting clinician. The oral report in addition to the computer provided e-record will also have its impact on clinician's judgment. To summarize, temporal and timing information, peripheral awareness of other clinicians, physicians and caregivers, and resource availabilities (including both records and devices) all belong to contextual information. In this section, we discuss how to use this contextual information to coordinate among different groups and organizations, various physicians, clinicians, and caregivers in the hospital [2, 71–75].

AwarePhone and AwareMedia are two tools that can run on mobile platforms. Basic activities can be directly reported on personal smart phones or computers—for example, the location of patient, the procedures of patient, the procedure of operations, and the time in operation. AwarePhone is a mobile app that provides physician location, patient status, and operating room availability on the physician's smart phone. It also allows messages and call exchanges among physicians and nurses [1, 73–76]. In addition to the features AwarePhone has, AwareMedia enables a chat channel for surgical usage to broadcast important messages to nurses and other clinicians notifying an emergency such as blue code. These contextual information tools intend to help hospital personnel to efficiently locate resources and effectively improve cooperation. If AwarePhone and AwareMedia target mainly physicians, then the MobileWard is designed for nurse stations or caregivers' sphere [73–76]. The tool coordinates patients' identity, temporal events with nurse's spatioinformation in real time. This tool can display patient chart, medical record, and a list of surgical procedures on ambient displays when the nurse is approaching certain locations such as right in front of surgical room. Additional software support can also be applied to the health information systems (HIS) [9]. Here, context-aware tools can help locate relevant medical images for an upcoming surgery [73–76]. These tools usually automatically provide patient medical images based on the progression of the surgery. HIS with context-aware capability is also referred to as context-aware HIS (or CHIS). In addition to the context-awareness like the AwarePhone, the AwareMedia, and the MobileWard, the CHIS also employs context to personalized information displayed or presented on computers and displays through controlled data privacy.

However, privacy and accuracy are the two major concerns in context awareness in hospitals. For example, it is very difficult to predict or estimate the temporal-related information during a complicated surgical procedure. Therefore, it is not possible to provide highly confident information to physicians. In Section 3.4, we also pointed out that the amount of data to process is extremely large in the pervasive computing environment at hospitals where a large number of patients with voluminous medical data need to be monitored at the same time. How to selectively display information on a mobile screen for a highly mobile physical remains a challenging data mining and system implementation research topic. The second concern is the privacy problem with personal information during interactive exchanging of messages, calls, and displays for patients, nurses, and physicians. It reveals their location, temporal information, and status of surgical procedure in real time. We will further discuss data and information security issue in Section 3.6.

3.6 DATA AND INFORMATION SECURITY

Recent years have witnessed a rapid growth of wearable wired and wireless devices for monitoring and retrieving information from HIS in a versatile manner [5, 77, 78]. Because of wearable devices' moderate cost and their ability to continuously collect data, healthcare providers believe that these devices "carry the promise of drastically improving and expanding the quality of care across a wide variety of settings and segments of the population" [5]. Indeed, as discussed earlier, a number of clinics and hospitals have already demonstrated the potential of patient monitoring through early prototypes including various forms of systems with imaging sensors and embedded systems [79]. As most wearable devices and medical imaging instruments are known to be resource constrained, great amount of research efforts have been invested on how to reduce the signal acquisition complexity on these medical systems and how to enhance the energy efficiency of data communication [5, 77, 78, 80]. For that purpose, a major trend today is to utilize CS [45, 46, 81], a recent ground-breaking data sampling and reconstruction framework that unifies the traditional sampling and compression process for data acquisition. We also have introduced CS concept in Section 3.4. By leveraging the sparsity of the data, CS enables sub-Nyquist sampling and low-energy data reduction, making the technique especially attractive in healthcare sensing monitoring systems.

CS [45, 46, 81], has received lots of attention in signal processing community recently. Because it unifies the sampling and compression for data acquisition, CS has been found to be of great interest in a number of recent biomedical sensing applications [77–80, 82]. But those works consider neither the cloud assistance to handle large amount of data nor the importance of securing the healthcare data. These are indispensable design requirements in our system design, and we maintain the same sensing simplicity and low bandwidth cost at the sensor. A line of research loosely related to the proposed work is the study on the security and robustness of encryption via compressive sensing, by Orsdemir et al. [83], Rachlin et al. [84], and others therein. They investigate the secrecy of the linear measurement by assuming that the adversary has no knowledge of the sensing matrix. Their results suggest that by keeping the sensing matrix as secret, approaches of brute-force searching that try to recover original data can be computationally infeasible. A recent work [85] by Divekar et al. shows how to leverage CS to compress the storage of existing correlated image datasets. The cloud-assisted image recovery over encrypted samples in our system is also akin to secure computation outsourcing [86–88] (to list a few), which aims to protect both input and output privacy of the outsourced workloads. With garbled circuits [89] and fully homomorphic encryption (FHE), a general solution is shown feasible in theory by Gennaro et al. [86], where the computation is represented by an encrypted combinational boolean circuit that allows to be evaluated with encrypted private inputs. But it is not practical for the extremely large circuit size and huge computation cost of FHE operations. Besides general solutions, researchers have been working on customized mechanisms for securely outsourcing specialized computations, such as matrix multiplications [88], modular exponentiations [87], etc. Though elegant, they are not directly applicable in our system as they handle different computations.

However, the growing popularity of these systems has also revealed several grand challenges yet to be addressed satisfactorily: how to effectively process the ever-growing healthcare data and simultaneously protect patient's data privacy, while maintaining the low overhead at the sensors? In order to create high-quality images or simulate motions and activities, healthcare systems usually require continuous and routine monitoring to capture a deluge of information. For example, a transceiver on a body sensor network monitoring physiological signal of an individual can easily capture and send nearly 2.77 GB of raw data per day. When the sample rate is high, such number can even reach up to 31 GB/day [77]. To effectively manage and process the huge amount of information, a natural choice is to offload the data to cloud for its economic yet abundant computing resources. But security inevitably becomes a major concern. This is not only because the servers and their formed local cloud (just for hospitals and clinics) are an open environment operated by external third parties [90] but also because many healthcare data, for example, the medical images with diagnostic results for different patients, are privacy sensitive by its nature [91]. Thus, it is of critical importance to ensure that security be embedded in the healthcare monitoring system design from the very beginning so that we can better protect patient's data without sacrificing the usability and accessibility of the information. For a truly practical security design, it is essential to still maintain the low cost on data sensing, as current healthcare monitoring system does with the choice of compressive sensing. To address these challenges, we discuss a cutting-edge cloud-assisted healthcare monitoring system architecture with privacy assurance. This system leverages techniques from different domains and achieves the following benefits. For efficient data acquisition at sensor, we discuss the system under the CS framework. Although it has been exploited recently in a number of biomedical sensors [77, 78, 80, 82], none of these works has ever considered the increasing importance of securing the healthcare data, especially in a local cloud-assisted monitoring system. For privacy protection, acquired sensitive samples never leave sensor in unprotected form. Such protected samples are later sent from sensors directly to cloud, which functions as a central hub responsible for storage, processing, and disseminating reconstructed data to receivers. The whole process is privacy assured such that the information from neither the samples nor the underlying data content will be revealed.

One key design consideration among others is to minimize the cost of sensors, in particular the communication cost, when generating and transmitting the protected samples to cloud. This stringent requirement practically excludes the applicability of existing techniques in the context of FHE [86] due to the hugely enlarged ciphertext and operational cost. Simply adopting ordinary symmetric encryption techniques over samples is also not a viable choice, as it in essence prevents the cloud from performing any meaningful operation of the protected samples, let alone data reconstruction. In our design, we leverage the fact that the CS data recovery can be achieved by solving a formulated linear programming (LP) problem, and investigate a secure transformation mechanism that randomly maps the original data recovery problem into a random one so as to hide the information of the underlying data. Such a random mapping design has the benefit of maintaining the same bandwidth cost at the sensor as current mechanism does without security consideration. In addition, it brings in

considerable computational savings at the receiver side without introducing extra computational complexity at the cloud. For the rest of this section, we start with the architecture design for the case of sparse data, which is the typical scenario for compressive sensing. Then we show its natural extension to the general data, which allows meaningful tradeoffs between efficiency and accuracy. In order to be truly powerful and robust, we also study how to make the design applicable to real-world scenarios where data samples are tampered with noise. This system is one of the first privacy-assured healthcare monitoring system with simplified data acquisition, secure cloud-assisted data reconstruction, and local resource savings. The system supports sparse data, general data, and data tampered with noise, under various application contexts. All cases support the aforementioned design features satisfactorily. Thorough security analysis and results show that the type of designs can indeed achieve security, efficiency, effectiveness, and resource savings, simultaneously. The service model envisioned by the cloud-assisted healthcare monitoring design can be illustrated in Figure 3.13.

To effectively manage and process the huge amount of sensed data, the sensor in our system will directly offload the acquired data samples to the local cloud. Those samples are all in protected form for privacy protection. Cloud with its abundant resources is responsible to provide various privacy-assured data services for receivers, such as on-demand data recovery, data retrieval, and others. Here the receiver might be a healthcare workstation operated by a physician in a hospital. In the following, we will use medical image reconstruction from compressed samples as a concrete application to demonstrate the system design. Under such an architecture, cloud is responsible for image recovery from the received samples. The sensor first acquires compressed samples from some sparse/compressible signal. Instead of directly sending original samples, the sensor first encrypts the original samples with a secret key and sends the encrypted samples to the local cloud. For healthcare monitoring, the aforementioned process is assumed to be in continuous and routine fashion, and the samples will be accumulated at the local cloud. Upon receiving a recovery request

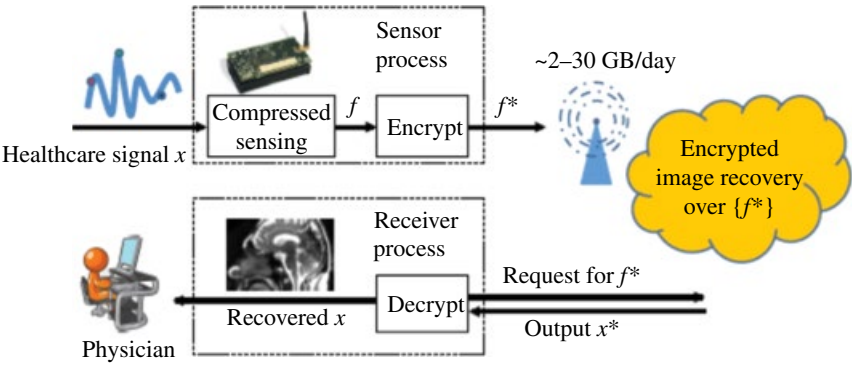


FIGURE 3.13 Wireless medical devices are connected to local cloud and continuously transmit data in real time [92].

with associated metadata from the receiver, the cloud recovers an output over some specified encrypted samples and hands it back to receiver. Finally, the receiver decrypts using the secret key. Because encrypted original samples and the original samples are with the same size, the security protection does not bring in any extra bandwidth cost at the sensor.

Due to the assistance of cloud, the local computation at the receiver is expected to be considerably reduced, and the overall image recovery efficiency improved. We assume an "honest-but-curious" cloud, which faithfully follows the designated protocol but may voluntarily infer data content for various purposes. Thus, any data leaving the data sensor/receiver have to be encrypted. The design goals consist of the following: (i) The healthcare monitoring system should be effective and reliable such that it correctly supports the cloud-assisted data recovery, while protecting both the private data samples and the content of the recovered data from the cloud; (ii) The system should maintain low local cost, especially the communication cost for acquired sample transmission, on the resource-constrained sensor. It should also keep the local computation at receiver substantially less than the service designs without cloud assistance; and (iii) The system should be designed to be compatible with other important data services, like content-based data/image retrieval, while providing possible extensible service interfaces. To instantiate a privacy-aware cloud-assisted healthcare system, we have to ensure acquired samples never leave the sensor in unprotected form. For energy efficiency, we have to maintain the low cost, especially the bandwidth cost for the sample acquisition and transmission, at the sensor. The protected samples should support processing for healthcare image recovery as needed. Moreover, the images recovered at cloud should still be in a protected form.

Due to different application contexts, the data to be sensed in healthcare systems can be of various types, including sparse data, general data, and data tampered with noise. Because the secret keying materials and sampling matrices can all be generated by some random seeds and thus easily shared, it allows us to use a fresh set of secret for every image to be captured and recovered. In other words, we can use independent random keys for different images. Recall that we also assume the orthonormal basis is shared properly between the sensor and the receiver. For different applications, the choice of this basis can be either data dependent, like Karhunen–Loeve basis, or data independent, like wavelet basis, both of which can be prescribed without much extra overhead.

In the real-world applications, there are many scenarios where healthcare data sources are not exactly sparse. A natural question would be the following: How to extend the application to those nonsparse data? To answer this question, we can use sparse data to well approximate general data [93], as long as discarding its small coefficients does not lose much information. Handling general data strikes a balance between efficiency and accuracy. If the general data are nearly sparse, the discussed system will provide good approximation. If the original general data are not quite compressible, this design will still recover the image at its best, by reconstruction from its sparse approximation. The efficiency and security of the healthcare monitoring system remains, but the quality of the recovered image might be downgraded.

The data collected from sensors in healthcare systems are not always of high quality. They could be tampered due to the errors in transmission channel, the noise brought by the imperfect sensing devices, etc. As healthcare staffs rely on these systems for accurate diagnosis, it is thus imperative for the system design to robustly handle nonideal scenarios and still provide acceptable quality of recovered images. We use oversamples to compensate the error corruption in the measurements, and this is the price we have to pay even in the plaintext scenario without security consideration. Because we rely on the local cloud to recover an error, the data flow will be slightly different from the case of sparse data and general data. Specifically, for each sampling, the sensor first acquires the oversampled linear measurements, encrypts them via some standard symmetric key encryption, for example, advanced encryption standard (AES), and then sends the encrypted samples to the local cloud. The encryption key can be generated and shared in the same fashion as the previous cases by using random coins. The same bandwidth overhead at the sensor side is still ensured as before while the samples get protected. For the data recovery, the receiver first requests the protected samples from the cloud. Figure 3.14(a1) and (a2) give two examples of the simulated signal source. The recovered image after the transformation and receiver decryption is shown in Figure 3.14(c1) and (c2). For good experiment results, we use relatively large number of linear measurements for each block recovery, whereas the sparsity level of many blocks is around a few dozen. This follows the "four-to-one" rule suggested by [93], that is, the number of samples is roughly 4 times the sparsity level of the targeted signals. For comparison purposes, recovered images using $m = 128; 192; 256$ measurements are shown in Figure 3.15. Figure 3.14(b1) and (b2) shows the recovered image before receiver decryption,

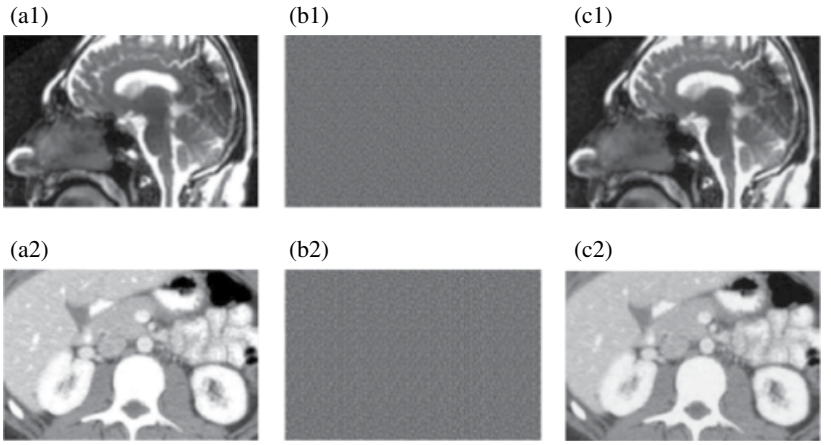


FIGURE 3.14 Demonstration of privacy assurance and effectiveness for the case of compressible data: (a1)(a2): simulated signal source; (b1)(b2) reconstruction via encrypted data; and (c1)(c2) reconstruction at receiver after decryption [92].

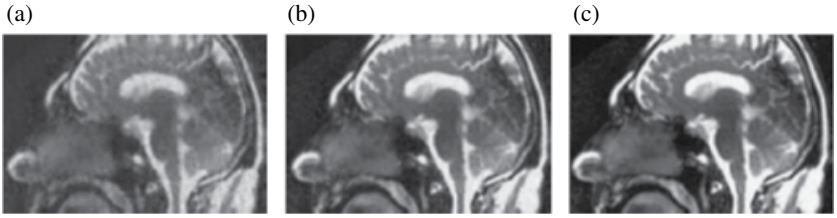


FIGURE 3.15 Recovered images via different measurements m [92]. (a) $m=128$, (b) $m=192$, and (c) $m=256$.

which based on the blinded vector is indistinguishable from a randomly generated image with only noise background, because of the random affine mapping.

3.7 CONCLUSION

This chapter introduces the general framework of pervasive computing in hospitals. Wearable medical devices, servers and local cloud, and their functionality and designs are discussed. Designs of wearable devices and the applications of these devices are exploited to reveal the pervasive computing mechanism at hospitals and clinics. We also introduced several popular contextual information-based software to demonstrate the cutting-edge temporal and spatiocoordinated resource management and personnel collaboration using advanced digital and mobile technologies. Accuracy and privacy are two major challenges for pervasive computing. The first one is also related to the hardware design and data volume. It is important to understand how to effectively reduce the amount of data and yet still keep core information. Privacy-aware cloud-assisted healthcare monitoring system via CS is the state-of-the-art technique that intends to secure data with compression. The design exploits techniques from different domains and achieves the following novel benefits. The sensor can utilize the framework of CS to consolidate the sampling and compression via only linear measurements. The random mapping-based protection ensures no sensitive samples would leave the sensor in unprotected form. Such a security approach also minimizes the communication cost for sensor data acquisition and transmission. On the receiver side, the cloud-assisted image recovery over encrypted samples provides great computational savings, yet without revealing either the received compressed samples or the content of the recovered underlying image.

ACKNOWLEDGMENT

We would like to thank Dr. Richard Lane, Prof. Cong Wang, Dr. Bijan Najafi, Mr. Dung Nguyen, and Mr. Kemeng Chen for their contributions to this chapter.

REFERENCES

[1] M. Tentori, G. Hayes, and M. Reddy, Pervasive computing for hospital, chronic, and preventive care, Foundations and Trends in Human-Computer Interaction, 5(1), 1–95, (2011).

[2] J. E. Bardram and C. Bossen, Moving to get aHead: Local mobility and collaborative work, in Proceedings of the Eighth International Conferences on Computer Supported Cooperative Work, September 14–18, 2003, Helsinki, Finland. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003, pp. 355–374.

[3] J. E. Bardram and C. Bossen, Mobility work: The spatial dimension of collaboration at a hospital, in International Conferences on Computer Supported Cooperative Work, April 1, 2005, Paris, France. Copenhagen, Denmark: Kluwer Academic Publishers, 2005, pp. 131–160.

[4] C. Bossen, The parameters of common information spaces: The heterogeneity of cooperative work at a hospital ward, in Computer Supported Cooperative Work, New Orleans, LA, November 16–20, 2002, Copenhagen, Denmark: Kluwer Academic Publishers, 2002, pp. 176–185.

[5] J. Ko, C. Lu, M. Srivastava, J. Stankovic, A. Terzis, and M. Welsh, Wireless sensor networks for healthcare, Proceedings of the IEEE, 98(11), 1947–1960, (2010).

[6] O. Chipara, C. Lu, T. C. Bailey, and G.-C. Roman, BReliable patient monitoring: A clinical study in a step-down hospital unit, Department of Computer Science and Engineering, Washington University St. Louis, St. Louis, MO, Technical Report WUCSE-2009-82, December 2009.

[7] J. Ko, J. Lim, Y. Chen, R. Musaloiu-E, A. Terzis, G. Masson, T. Gao, W. Destler, L. Selavo, and R. Dutton, BMEDiSN: Medical emergency detection in sensor networks, ACM Transactions on Embedded Computing Systems, 10(1), 11:1–11:29, (2010), article 11.

[8] The Reaction Platform. http://www.reactionproject.eu/articles.php?article_id=3 (accessed January 2, 2016).

[9] Health Information Systems (HIS). Pacific Health Information Network. http://phinnetwork.org/Resources/HIS.aspx (accessed January 2, 2016).

[10] J. Jensen, D. Chang, and E. Lee, A model-based design methodology for cyber-physical systems, in Proceedings of the 1st IEEE Workshop on Design, Modeling, and Evaluation of Cyber-Physical System (CyPhy), Wireless Communications and Mobile Computing Conference (IWCMC), Istanbul, Turkey, July 6, 2011, pp. 1666–1671.

[11] C. Brooks, C. Cheng, T. Feng, E. A. Lee, and R. von Hanxleden, Model engineering using multimodeling, in Proceedings of the 1st International Workshop on Model Co-Evolution and Consistency Management (MCCM 08), Center for Hybrid and Embedded Software Systems (CHESS) at UC Berkeley, April 16, 2008.

[12] K. Balasubramanian, A. Gokhale, G. Karsai, J. Sztipanovits, and S. Neema, Developing applications using model-driven design environments, IEEE Computer, 39(2), 33, (2006).

[13] G. Karsai, J. Sztipanovits, A. Ledeczi, and T. Bapty, Model-integrated development of embedded software, Proceedings of the IEEE, 91(1), 145–164, (2003).

[14] J. Eker, J. Janneck, E. A. Lee, J. Liu, X. Liu, J. Ludvig, S. Neuendorffer, S. Sachs, and Y. Xiong, Taming heterogeneity: The Ptolemy approach, Proceedings of the IEEE, 91(1), 127–144 (2003).

[15] N. Cutland, Computability: An Introduction to Recursive Function Theory. Cambridge, MA: Cambridge University Press, 1997, pp. 100–112, 149–156.

[16] E. A. Lee, Computing needs time, ACM Communications, 52(5), 70–79, (2009).

[17] J. Eidson, E. A. Lee, S. Matic, S. A. Seshia, and J. Zou, Time-centric models for designing embedded cyber-physical systems, University of California, Berkeley, Technical Memorandum, UCB/EECS-2009-135, October 2009.

[18] J. Tyson, How Encryption Works, HowStuffWorks, 2001. http://computer.howstuffworks.com/encryption3.htm (accessed February 15, 2016).

[19] J. Vollbrecht, The Beginnings and History of RADIUS, 2006. http://www.interlinknetworks.com/app_notes/History%20of%20RADIUS.pdf (accessed January 2, 2016).

[20] B. Posey, SolutionBase: RADIUS Deployment Scenarios. TechRepublic. Retrieved April 15, 2009, (2006).

[21] Cisco Engineers. How Does RADIUS Work?. Retrieved April 15, 2009 (2006).

[22] E. L. Brown, 802.1X Port-Based Authentication. New York: Taylor & Francis, 2006. p. 17.

[23] WHO. http://www.who.int/medical_devices/full_deffinition/en/ (accessed January 2, 2016).

[24] D. Nguyen, D. Ma, and J. Roveda, Recent design development in molecular imaging for breast cancer detection using nanometer CMOS based sensors. Journal of Oncology. 2012, 1–6, 2012.

[25] Zephyr Life, Technical application node, Nieuws_Zephyr en ECG.pdf

[26] D. Nguyen, D. Ma, and J. Roveda, Recent design development in molecular imaging for breast cancer detection using nanometer CMOS based sensors. Journal of Oncology. 2012, 1–6, 2012.

[27] N. Arzeno, Z. Deng, and C. Poon, Analysis of first-derivative based QRS detection algorithm, IEEE Transactions on Bio-Medical Engineering, 55(2), 478–484, (2009).

[28] V. Afonso, ECG QRS Detection, 1993. http://www.ejwoo.com/uploads/2/5/4/0/25400517/ecg_qrs_detection.pdf (accessed February 19, 2016).

[29] S. N. Narouze (ed.), Atlas of Ultrasound-Guided Procedures in Interventional Pain Management. New York, NY: Springer Science+Business Media, LLC, 2011.

[30] Ultrasound system. Texas Instruments. http://www.ti.com/solution/ultrasound_system (accessed January 2, 2016).

[31] Field II, Calculation of B-mode image of synthetic kidney, 2012. http://field-ii.dk/?examples/kidney_example/kidney_example.html (accessed January 2, 2016).

[32] J. Scampini. Overview of Ultrasound Imaging Systems and the Electrical Components Required for Main Subfunctions. May 10, 2010. http://www.maximintegrated.com/app-notes/index.mvp/id/4696 (accessed January 2, 2016).

[33] Field II. Field II Ultrasound Simulation Program, 2016. http://field-ii.dk/ (accessed January 2, 2016).

[34] European Hospital, Digital Mammography and Tomosynthesis, December 31, 2007, http://www.european-hospital.com/en/article/3197-Digital_mammography_and_tomosynthesis.html (accessed January 2, 2016).

[35] Department of Radiology. UC Davis Health System: Department of Radiology—Breast CT Scanner for Earlier Cancer Detection. 2006. http://www.ucdmc.ucdavis.edu/radiology/research/boone3.html (accessed January 2, 2016).

[36] A. O'Connell, D. L. Conover and Y. Zhang, Cone-beam CT for breast imaging: Radiation dose, breast coverage, and image quality, AJR: American Journal of Roentgenology, 195,

496–509, (2010), http://www.ajronline.org/content/195/2/496.full.pdf+html (accessed January 2, 2016).

[37] S. S. J. Feng and I. Sechopoulos, Clinical digital breast tomosynthesis system: Dosimetric characterization, Radiology, 263(1), 35–42, (2012), http://radiology.rsna.org/content/early/2012/02/09/radiol.11111789.full (accessed January 2, 2016).

[38] O. Yadid-Pecht, Wide dynamic range sensors, Optical Engineering, 38(10) 1650–1660, (1999).

[39] R. Gonzalez and R. Woods, Digital Image Processing. Reading, MA: Addison Wesley Publishing Company, 3rd Edition.

[40] E. Gamal, High dynamic range image sensors, in IEEE International Solid State Circuit Conference Tutorial # 3, San Francisco, CA, February 7, 2002, pp. 1–62. http://cafe.stanford.edu/~abbas/group/papers_and_pub/isscc02_tutorial.pdf (accessed February 19, 2016).

[41] B. Dally, Low-Power Architecture, Keynote, ISSCC, 2005.

[42] M. Pedram, Low Power Design Methodologies and Techniques: An Overview, Invited Talk at Low Power Design Conference, 2012, pp. 1–34. http://www.mpedram.com/Papers/LPD-talk.pdf (accessed February 19, 2016).

[43] P. Hofstee, Power efficient processor architecture and the cell processor, in Proceedings of the 11th International Symposium on High-Performance Computer Architecture, San Francisco, CA, February 12–16, 2006, IEEE Computer Society, Washington, DC, pp. 258–262.

[44] Texas Instruments, 16-Channel, 24-Bit Analog to Digital Converter, Texas Instruments Data Sheet on ADS 1258 for years 2005–2011. http://www.ti.com/lit/ds/symlink/ads1258.pdf (accessed January 2, 2016).

[45] D. Donoho, Compressive sensing, IEEE Transactions on Information Theory, 52(4) 1289–1306, (2006).

[46] E. Candes, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory, 52(2), 489–509, (2006).

[47] E. J. Candès, J. K. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics, 59(8), 1207–1223, (2006).

[48] E. Candès and J. Romberg, Encoding the l_p ball from limited measurements, in Proceedings of the IEEE Data Compression Conference, Snowbird, UT, March 28–30, 2006, IEEE, pp. 33–42.

[49] R. Boufounos and R. G. Baraniuk, Quantization of sparse representations, in Proceedings of the Data Compression Conference, Snowbird, UT, March 27–29, 2008, IEEE Computer Society, p. 378.

[50] P. Boufounos and R. G. Baraniuk, Sigma delta quantization for compressive sensing, in Proceedings of the Wavelets XII in SPIE International Symposium on Optical Science and Technology, 6701, 6701–6704, August 26–29, 2007, San Diego, CA.

[51] P. Boufounos and R. G. Baraniuk, 1-Bit compressive sensing, in Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, March 19–21, 2008, pp. 16–21.

[52] V. Goyal, A. Fletcher, and S. Rangan, Compressive sampling and lossy compression, IEEE Signal Processing Magazine, 25, 48–56, (2008).

[53] L. Jacques, D. Hammond, and M. Jalal Fadili, Dequantizing compressed sensing: When oversampling and non-Gaussian constraints combine. arXiv:0902.2367, 2009.

[54] W. Dai and O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Transactions on Information Theory, 55(5), 2230–2249, (2009).

[55] D. Needell and J. A. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Applied and Computational Harmonic Analysis, 26(3), 301–321, (2009).

[56] Rice University Digital Signal Processing Group. Compressive imaging: a new single-pixel camera. http://dsp.rice.edu/cscamera (accessed January 2, 2016).

[57] M. Lin, Analogue to Information System Based on PLL-Based Frequency Synthesizers with Fast Locking Schemes, Ph.D. thesis, The University of Edinburgh, February 2010.

[58] N. G. Prelcic, State of the art of analog to information converters, keynote presentation, 2010.

[59] M. Mishali, R. Hilgendorf, E. Shoshan, I. Rivikin, and Y. C. Eldar, Generic sensing hardware and real-time reconstruction for structured analog signals, in Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Rio de Janeiro, May 15–18, 2011.

[60] M. Mishali, Y. C. Eldar, and A. Elron, Xampling: Signal Acquisition and Processing in Union of Subspaces, CCIT Report no. 747, EE Dept., Technion; arXiv.org 0911.0519, October 2009.

[61] Y. C. Eldar, Compressed sensing of analog signals in shift-invariant spaces, IEEE Transactions on Signal Processing, 57(8), 2986–2997, (2009).

[62] M. Vetterli, P. Marziliano, and T. Blu, Sampling signals with finite rate of innovation, IEEE Transactions on Signal Processing, 50(6) 1417–1428, (2002).

[63] A. Bandyopadhyay, J. Lee, R. W. Robucci, and P. Hasler, Matia: A programmable 80 uW/frame CMOS block matrix transform imager architecture, IEEE Journal of Solid-State Circuits, 41(3), 663–672, (2006).

[64] A. Bandyopadhyay and P. Hasler, A fully programmable CMOS block matrix transform imager architecture, in Proceedings of the IEEE 2003 Custom Integrated Circuits Conference (CICC), San Jose, CA, September 21–24, 2003, pp. 2233–2236.

[65] A. Bandyopadhyay, P. Hasler, and D. Anderson, A CMOS floating-gate matrix transform imager, IEEE Sensors, 5(3), 455–462, (2005).

[66] P. Hasler, A. Bandyopadhyay, and D. V. Anderson, High fill-factor imagers for neuromorphic processing enabled by floating gates, EURASIP Journal on Advances in Signal Processing, (7), 676–689, (2003).

[67] V. Srinivasan, G. J. Serrano, J. Gray, and P. Hasler, A precision CMOS amplifier using floating-gates for offset cancellation, in Proceedings of the IEEE 2005 Custom Integrated Circuits Conference (CICC), San Jose, CA, September 21, 2005, pp. 734–737.

[68] J. Favela, M. Tentori, D. Segura, and G. Berzunza, Adaptive awareness of hospital patient information through multiple sentient displays, International Journal of Ambient Computing and Intelligence, 1(1), 27–38, (2009).

[69] D. Segura, J. Favela, and M. Tentori, Sentient displays in support of hospital work, in Proceedings of the 3rd Symposium of Ubiquitous Computing and Ambient Intelligence, Salamanca, Spain, October 22–26, 2008, pp. 103–111.

[70] M. Munoz, M. D. Rodriguez, J. Favela, A. I. Martinez-Garcia, and V. M. Gonzalez, Context-aware mobile communication in hospitals, IEEE Computer, 36(9), 38–46, (2003).

[71] T. Chin, Untapped power: A physician's handheld, in *American Medical News*, January 17, 2005. Available at http://www.ama-assn.org/amednews/site/free/bisa0117.htm (accessed January 2, 2016).

[72] M. Reddy and P. Dourish, A Finger on the Pulse: Temporal Rhythms and Information Seeking in Medical Work, in Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, ACM Press, New Orleans, LA, November 16–20, 2002, pp. 344–353.

[73] M. Reddy, P. Dourish, and W. Pratt, Coordinating heterogeneous work: Information and representation in medical care, in Proceedings of the Seventh European Conferences on Computer-Supported Cooperative Work, Bonn, Germany, September 16–20, 2001, Springer, pp. 239–258.

[74] J. Bardram and T. Hansen, The AWARE architecture: Supporting context mediated social awareness in mobile cooperation, in ACM Conference on Computer Supported Cooperative Work (CSCW 2004), Chicago, IL, November 6–10, 2004, ACM Press, pp. 192–201.

[75] S. Agarwal, A. Joshi, T. Finin, Y. Yesha, and T. Ganous, A pervasive computing system for the operating room of the future, Mobile Networks and Applications, 12(2–3), 215–228, (2007).

[76] J. E. Bardram and N. Nørskov, A context-aware patient safety system for the operating room, in UbiComp '08: Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, September 21–24, 2008, ACM Press, pp. 272–281.

[77] M. Shoaib and H. Garudadri, Digital pacer detection in diagnostic grade ECG, in Proceedings of the 13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom), Columbia, MO, June 13–15, 2011, pp. 326–331.

[78] M. Shoaib, N. Jha, and N. Verma, A compressed-domain processor for seizure detection to simultaneously reduce computation and communication energy, in Proceedings of IEEE Custom Integrated Circuits Conference, San Jose, CA, March 9–12, 2012, IEEE, pp. 1–4.

[79] J. Ko, J. Lim, Y. Chen, R. Musaloiu-E, A. Terzis, G. Masson, T. Gao, W. Destler, L. Selavo, and R. Dutton, BMEDiSN: Medical emergency detection in sensor networks, ACM Transactions on Embedded Computing Systems, 10(1), 11:1–11:29, (2010).

[80] W. Xu, M. Zhang, A. Sawchuk, and M. Sarrafzadeh, Corecognition of human activity and sensor location via compressed sensing in wearable body sensor networks, in Proceedings of the IEEE Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN), London, May 9–12, 2012, pp. 124–129.

[81] E. Candes and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory, 52(12), 5406–5425, (2006).

[82] F. Chen, A. Chandrakasan, and V. Stojanovic, A signal-agnostic compressed sensing acquisition system for wireless and implantable sensors, in Proceedings of the IEEE 32nd Annual Custom Integrated Circuits Conference on The Showcase for Circuit Design in the Heart of Silicon Valley, San Jose, September 19–22, 2010, IEEE, pp. 1–4.

[83] A. Orsdemir, H. Altun, G. Sharma, and M. Bocko, On the security and robustness of encryption via compressed sensing, in Proceedings of the Assuring Mission Success 2008 IEEE Military Communications Conference (MILCOM), Washington, DC, November 17–19, 2008, pp. 1–7.

[84] Y. Rachlin and D. Baron, The secrecy of compressed sensing measurements, in Proceedings of the IEEE 46th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 23–26, 2008, IEEE, pp. 813–817.

[85] A. Divekar and O. Ersoy, Compact storage of correlated data for content based retrieval, in Proceedings of the 43rd Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 1–4, 2009, pp. 109–112.

[86] R. Gennaro, C. Gentry, and B. Parno, Non-interactive verifiable computing: Outsourcing computation to untrusted workers, in Proceedings of the 30th Annual Cryptology Conference, Santa Barbara, CA, August 15–19, 2010, Springer, pp. 465–482.

[87] S. Hohenberger and A. Lysyanskaya, How to securely outsource cryptographic computations, in Proceedings of the Theory of Cryptography, Second Theory of Cryptography Conference (TCC), Cambridge, MA, February 10–12, 2005, pp. 264–282.

[88] M. Atallah and K. Frikken, Securely outsourcing linear algebra computations, in Proceedings of the ASIACCS, Beijing, China, April 13, 2010, ACM, New York, pp. 48–59.

[89] A. Yao, Protocols for secure computations (extended abstract), in Proceedings of the IEEE 23rd Symposium on Foundations of Computer Science (FOCS), Chicago, IL, November 3–5, 1982, IEEE, pp. 160–164.

[90] European Network and Information Security Agency, Cloud computing risk assessment. Online at http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment (accessed January 2, 2016), (2009).

[91] 104th United States Congress, Health Insurance Portability and Accountability Act of 1996 (HIPPA). Online at http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html (accessed January 2, 2016), (1996).

[92] C. Wang, B. Zhang, K. Ren, C. Chen, Z. Xu, and J. M. Roveda, A privacy-aware cloud-assisted healthcare monitoring system via compressive sensing, Proceedings of the IEEE Infocom, 2130–2138, (2014).

[93] E. Candes and M. Wakin, An introduction to compressive sampling, IEEE Signal Processing Magazine, 25(2), 21–30, (2008).

4

DIAGNOSTIC IMPROVEMENTS: TREATMENT AND CARE

Xiaojun Xian

Center for Bioelectronics and Biosensors, The Biodesign Institute, Arizona State University, Phoenix, AZ, USA

4.1 INTRODUCTION

The traditional healthcare system faces new challenges in providing early stage diagnosis and personalized treatment for a large population of patients with chronic diseases [1, 2]. In the current healthcare system, diagnoses are performed in hospitals and clinics with the help of doctors or professionals by using specialized equipment. This is not suitable for patients with chronic diseases, such as certain neurological, respiratory, and cardiovascular diseases requiring constant monitoring of the patient over a long period of time. The frequent visiting of hospitals and clinics imposes additional financial and psychological burdens on patients. Another disadvantage is that the hospital-based diagnosis can only provide a snapshot of the patient's condition; while the patient's condition is actually a dynamic process requiring continuous monitoring. The continuous monitoring of patients provides information for effective diagnosis and treatment. The optimal diagnosis should be carried out when patients are in free-living conditions, under which the physiological parameters are more reliable to reflect the real health status. For instance, some researchers have shown that the blood pressure of patient tends to be much higher than normal when the patient is visiting their physicians [3], which may lead to misdiagnosis of hypertension.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

Thus, a new type of diagnosis technology with the feature of continuous monitoring of patients under free-living conditions is more favored and would greatly benefit the patient's suffering from chronic diseases. Continuous monitoring of a patient under free-living condition allows for long-term and overall profiling of key physiological parameters, from which the overall direction, magnitude, trend, and patterns can be obtained to provide better understanding and treatments. One example is that some alarm algorithms can be set up based on continuous monitoring data to alert the patient to take immediate actions to reduce the health risks. Some disease triggers for the patient can be identified by checking the correlation between environmental factors and the symptoms of the disease, such as an asthma attack. Medication dosages could also be dynamically adjusted and recommended to the patient based on the monitoring of certain biomarkers. Furthermore, the concept of closed-loop (sense/release) system can also be achieved by using a continuous monitoring system as a feedback loop with an automatic medication delivery system, such as a drug-delivery pump (like insulin pump), as a real-time treatment loop [4].

The rapid development of sensing technology, computing technology, and wireless communication technology has provided a solid technology foundation for the emergence of this new type of diagnosis technique as mentioned before. According to Moore's law, the number of transistors in the integrated circuit approximately doubles in every 2 years. The technology advance in semiconductor industry has provided strong driving force for the computer development by providing more powerful processing abilities, miniaturized size, and low cost, which make pervasive computing available in daily life. The fast growth of PDA technologies, the emergence of miniaturized sensing techniques, the advance of wireless communication technologies, and the widespread use of the Internet make pervasive health care available and attractive for the patients suffering from chronic diseases.

The key step for providing pervasive healthcare service to patients is the development of wireless BSN platforms for continuous human physiological parameter monitoring. In order to achieve continuous and reliable monitoring of patients in free-living conditions, the BSN system needs to have the following features: (i) simultaneous and continuous monitoring of multiple physiological parameters, (ii) nonobstructive design to make the sensors invisible to the patient, (iii) ubiquitous computing ability at each sensor node, (iv) wireless communication capability with the gateway, and (v) an effective interaction with healthcare professionals. Wearable electronics that implement miniaturized, flexible, or printable sensors in textiles and wearable gadgets hence are ideal candidates for building the BSN.

Besides the technical innovation of pervasive monitoring systems based on BSN, which makes it possible to obtain the right information at the right time and at the right place, a new concept of "personalized health care" can also be derived. This new concept requires the monitoring system to be customized to specifically fulfill the healthcare needs of each individual [5]. On the one hand, the BSN could be customized for each patient. Depending on the health condition of the patient, the relevant sensor nodes can be selected to build the sensor network. For example, the BSN for an asthma patient should be different from the BSN for a diabetic patient. On the other hand, considering the differences among patients with the same disease, the

algorithms in the system should be able to detect and identify the health baseline of the individual and apply a smart decision-making mechanism to trigger alarms and promote necessary actions for the treatment. Additionally, the BSN should allow the patient to effectively and privately communicate with healthcare professionals for an early diagnosis, early intervention, and early treatment.

The era of pervasive computing is approaching and it will inevitably change our life. The implementation of pervasive computing to the healthcare system will generate a new disease diagnosis and treatment model through constant monitoring of the patient over a long period of time, possibly even over the entire life of the patient.

4.2 SYSTEM DESIGN

The development of constant patient monitoring systems requires interdisciplinary teamwork including, but not limited to, expertise from bioengineering, chemical engineering, mechanical engineering, electrical engineering, software engineering, and user experience enhancement. The architecture should be very well defined so that efforts from different experts can be efficiently organized and integrated.

Until now, numerous layouts of constant patient monitoring system architectures have been proposed in the literature, with some prototype models demonstrated. There are two key components in the pervasive health monitoring system: the BSN and the MBU. The BSN consists of a group of sensor devices incorporated together for monitoring physiological parameters. Usually, these sensor devices are miniaturized units with the ability to continuously sense and transmit wireless data. The MBU includes three basic capabilities: wireless communication with sensor networks, health data processing, and communication with the medical center. The MBU serves as a gateway to bridge personal BSN and healthcare providers [6]. The typical architecture of constant patient monitoring system is illustrated in Figure 4.1.

In addition to the technical concerns, the architecture should also consider the medical service infrastructure, mainly on its standardization, interoperability, and extensibility. Triantafyllidis et al. proposed an open and reconfigurable BSN for pervasive health monitoring, with particular emphasis in its easy extension; additional sensors and functionality are added by incorporating embedded intelligent mechanisms [7]. In their design, each sensor node in the BSN will not only serve as the data source but also will have the capability of carrying out appropriate data processing algorithms to identify normal and problematic situations when monitoring the patient and to transmit the health data only when it is necessary. This approach emphasizes sensor node data processing capability and intrasensor intelligence that are technically achievable; it also reduces the burden on communication by only selectively transmitting the necessary health data. There are three features in this proposed architecture: (i) The entire wireless sensor network can be well described since each sensor node can provide a standard self-description. (ii) A sensor node can be simply added and removed from the entire network through the plug-and-play mode to provide personal sensor network for each individual patient. (iii) According to the operation need, the sensor nodes can be reconfigurable in run-time to make the

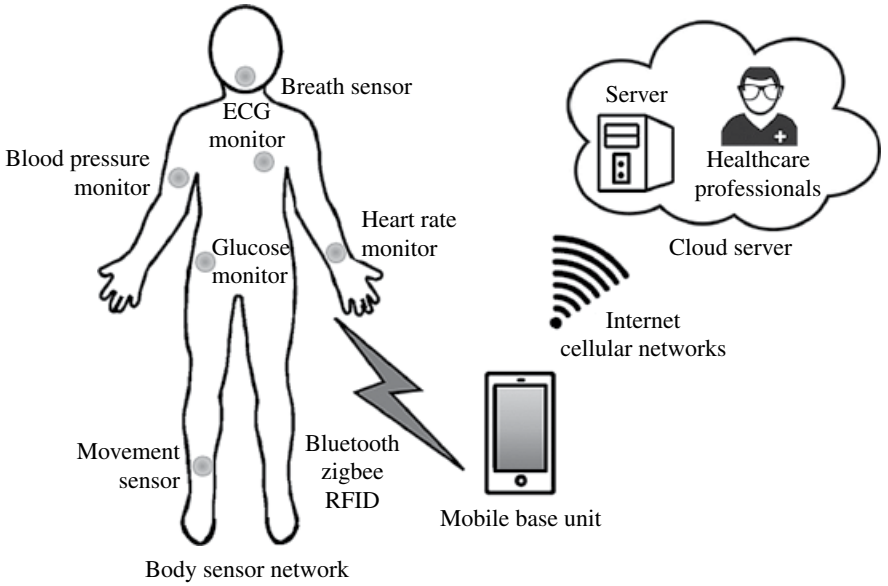


FIGURE 4.1 The typical architecture of constant patient monitoring system.

system more adaptive. Different from the wireless sensor networks, the MBU takes responsibilities for data aggregation, data handling, communication, as well as creating an adaptive-to-description user interface to assist the healthcare professionals analyze key monitoring parameters in line with the patient's personal profile.

4.3 BODY SENSOR NETWORK

The BSN refers to the human body monitoring platform composed of miniaturized sensors with the capability of data processing and transmission, which includes both wearable sensors and the implantable sensors. The wireless BSN is preferred to have on-body computational capability to provide better pervasive monitoring. As the basic elements of the BSN, each sensor node detects a specific physiological parameter of the body. When the data from each sensor node are integrated and assembled, a complete health profile of the patient can be obtained. Moreover, the BSN with continuous monitoring capability can provide a dynamic and real-time health profile of the patient. To achieve the functions mentioned earlier, the sensor node typically can be a stand-alone device with the following modules.

The sensing module is the core part of the sensor node. Since the sensor will be worn or carried by the patient in free-living conditions, only sensing technologies that can be realized as miniaturized units can be used to build the wireless BSN. These sensing technologies include, but are not limited, to the following: (i) optical detection methods, which monitor the reflection, absorbance, or transmission of optical signals to obtain the physiological parameters from the body, such as blood

oxygen sensors and heart rate monitors based on infrared optical signals; (ii) electrochemical sensing technologies, which determine the concentration of certain biomarkers in blood, sweat, saliva, urine, or even breath via specific electrochemical reaction of the targeting analyte on the electrodes surface by measuring the reaction-induced current in the electrochemical cell, such as the blood glucose meter, blood ketone meter, and breath nitric oxide (NO) monitor; (iii) impedance method measures the skin impedance between two electrodes, like electrocardiogram (ECG) recorder that incorporates the dry ECG electrodes with clothes for comfortable measuring of ECG signals; (iv) pressure sensors for blood pressure measurement; (v) accelerometer for movement and physical activity tracking, like pedometers; and (vi) global positioning system (GPS) for position tracking.

The battery provides power for the entire BSN. Ideally, each sensor node has its own battery since the voltage and power consumption between sensor nodes can vary. Another advantage of having a separate battery for each sensor node is that all the sensors in the BSN can work independently: in case of battery failure in one sensor node, which will not affect the operation of other sensor nodes. Batteries with high power capacity and miniaturized size are highly preferred in the sensor node design. A built-in rechargeable battery could be used in the sensor node since it lasts longer, and the user only needs to charge the battery when the power is low. Some energy harvesting technologies have also been proposed for the wearable sensors, which convert the energy from the body movements to electricity to drive the sensors. This is a clean way of generating electricity for the BSN, but this technology is still more conceptual than real. In addition to employing high-performance batteries, smart circuit design can also be used to save power consumption. Perhaps, some sensor nodes do not need continuous running. In this case, the microcontroller can automatically turn off the power when the sensor node is not actively running for a certain period of time. Furthermore, special protection needs to be considered in the design to make the monitoring system robust and waterproof so that it can endure different environment.

The microcontroller is the commander of the sensor node. It controls the active components in the sensor node to ensure that the entire system is functioning correctly and the true signal can be obtained. The microcontroller should have the basic data processing capability to handle the detected signals. Usually, raw optical or electrical signals need to be processed to achieve a better signal-to-noise ratio. Then the processed signal will be converted into the sensor response, which usually is the analyte concentration or certain physiological parameter. In some cases, the microcontroller should also be able to determine whether the obtained result is true or false, and whether it is normal or abnormal. The processed data are then transmitted wirelessly by the transmitter from the sensor node to the MBU.

The transmitter is a critical component of the sensor node, which handles the wireless data transmission from the BSN node to a local processing unit (LPU) or the MBU. This is an essential data collection step for the MBU from all of the sensor nodes in the BSN. The collected data will be further processed, analyzed, and integrated in the MBU and then sent to the health professional server via the Internet or mobile phone network. The data transmission between sensor node in the BSN and MBU can be either real time with no interruption, or periodical with interruption in

between, depending on the data collection methods and the types of the physiological parameters. One of the common transmitters used in the BSN sensor node design is the Bluetooth chip since it is of low cost and compatible with most PDAs.

In order to monitor the physiological parameters of the patients in free-living conditions, the BSN should be designed to be comfortable, unobtrusive, and secure. In this way, it will be transparent to the patient to avoid behavior changes. Wearable electronics are the ideal candidates for building the BSN, because the miniaturized sensors are usually hidden in clothing or wearable accessories, such as watches, glasses, belts, shoes, jewelry, headgears, arm wears, or other pieces. As a result, the sensor is essentially invisible and intangible to the patient. More details about the wearable sensors will be introduced and discussed in Section 4.5.

4.4 PORTABLE SENSORS

The goal of constant monitoring of the patient suggests the following requirements: (i) the monitoring should be performed without the help of healthcare professionals and (ii) the device has to be highly portable or preferably wearable so that the sensors can be with the patient at all times for immediate or continuous monitoring.

Several technical challenges need to be overcome in sensor node research and development to achieve the goal of constant monitoring. The first challenge is to find a strategy to build miniaturized sensor devices functioning as conventional bulky clinical equipment. This is not a simple scale-down process because the performance of the sensor, such as the sensitivity, selectivity, detection limit, and stability should not be sacrificed in the miniaturization. The adoption of new technologies and the creation of innovative integration methods are key factors for miniaturized sensor development. Further study on microelectromechanical systems/nanoelectromechanical systems (MEMS/NEMS) technologies [8] and new materials, such as intelligent textiles [9], would be of help to address this issue. The second challenge is the cost. High price will be a thick barrier preventing these technologies from benefiting patients. The last challenge is how to make the sensor easy to use. As mentioned earlier, the sensors will be used and maintained by the patients themselves. An easy-to-use operation procedure and intuitive design must be considered in the product development. Otherwise, the technical barrier can also deter patients away.

Wearable sensors are the ultimate goals for constant monitoring because they provide minimum obstruction to the patient's daily life. Unfortunately, the development of wearable sensors is still in the early stage and its social acceptance is still an important concern. Portable devices are compromised substitutes to meet the urgent need in this field. Additionally, the technologies accumulated in portable devices can provide a solid foundation for the wearable sensors design in the future.

Portable sensor device usually refers to a stand-alone and hand-held device, which can quantitatively measure the concentration of certain analyte. Generally, portable sensor devices have a wide range of sizes, from the size of a phone to the size of a basketball. For example, the size of Breezing metabolic tracker is 1.8 in. × 2.1 in. × 4.8 in., which can be easily held in hand for metabolism measurement (www.breezing.com).

Portable sensor devices have wide applications because of the following reasons: (i) Many physiological parameters of the patient are relatively stable for a certain period of time, for example, glucose. Thus, continuous monitoring is not necessary. The patients just need to check it a few times a day or even once a day, which means that a portable device will be more appropriate since the patient does not need to carry it constantly. (ii) Some biomarkers are not suitable for continuous monitoring, for example, biomarkers in breath, like NO for asthma. In this scenario, the point measurement rather than continuous measurement will make much more sense. (iii) Some measurements are invasive and may need to be carried out by portable devices.

The general design of portable design consists of the following modules: sampling and conditioning module, sample delivery module, detection module, signal processing and data saving module, and communication module.

The following portable devices are introduced in this section: (i) breath analyzers: NO devices and metabolic analyzer; (ii) blood analyzers: glucose meter and blood ketone monitor; (iii) others: blood pressure meter, blood oxygen meter, heart rate monitor, and breath rate monitor.

4.4.1 Breath Analyzers

Human breath is composed of thousands of compounds, many of which have been identified as biomarkers for certain diseases. The concentration of the biomarker in patient's breath can help with the diagnosis of certain type of disease and thus help with the treatment.

4.4.1.1 eNO Analyzer Exhaled NO (eNO) has been proven to be a biomarker of asthma, since the concentration of eNO is found to be higher among asthma patients [10]. This finding led to the development of noninvasive, reliable, and easy-to-use eNO analyzers for clinical applications. Such eNO analyzers can be used for the diagnosis of asthma, evaluation of the effectiveness of an anti-inflammatory treatment, and management of the dosage of anti-inflammatory medications. Furthermore, eNO could also be used in differential diagnosis. It is found that, in contrast to asthma patients, the eNO from patients with chronic obstructive pulmonary disease (COPD) or cystic fibrosis is not changed [11, 12].The chemiluminescence analyzer is the gold standard for gas-phase NO detection [13], but it is too bulky and expensive for personal use. A hand-held device for eNO detection has been developed to fulfill the needs for the intended clinical use and personal disease management. This device shows comparable precision and linearity with the widely accepted chemiluminescence technique. A product, called NIOX MINO, based on this technology has been available on the market to help asthma patients for home use [14]. Besides the function of saving data to the test card, this device also allows data to be wirelessly transmitted to other external devices through an IR port, such as a printer. This wireless data transmission capability could potentially be used to transmit the eNO concentration data to the MBU. In this way, the NIOX MINO can serve as a sensor node in the BSN system for asthma patient monitoring, though this is not a passive measurement and requires intensive involvement from the patient.

4.4.1.2 Metabolic Analyzer Resting metabolic rate (RMR) is a very important physiological parameter, which indicates the energy consumption of the human body. RMR is defined as the energy the body needs to maintain the functions of the vital organs such as the heart, lung, brain, muscles, nervous system, kidneys, etc. The monitoring of RMR is important for patients whose energy expenditure could undergo significant variation during the treatment. RMR can be used to guide the nutrition for patients with high risks for under- or overfeeding, or to assess the patients with obesity or thyroid disorders [15]. Besides medical applications, RMR has also been used for weight loss and fitness management. Some equations such as the renowned Harris–Benedict equation [16] or improved equations [17] have been widely used for RMR estimation. Since RMR depends on age, gender, genetics, and other attributes of the person, which vary widely from person to person, the estimated RMR from the equation can significantly differ from the person's true RMR value. Indirect calorimetry is the most well established approach for accurate assessment of RMR, which measures the RMR via the Weir equation [18] by assessing the oxygen-consumption rate and carbon dioxide production rate. Divergent from the traditional metabolic cart, which is bulky and expensive, several portable metabolic analyzers based on indirect calorimetry have been developed and brought to the market, such as BodyGem [19] and Breezing [20]. Breezing was designed to have the wireless data transmission capability through Bluetooth to personal smartphone, as shown in Figure 4.2. This feature makes it appropriate to serve as a sensor node in the BSN to monitor RMR for patients that need nutrition guidance.

Breath analysis for cancer diagnosis, including lung cancer [21] and breast cancer [22], has also been reported, and some point-of-care instruments, such as BreathLink™, are available on the market. More research is needed to shrink the size and reduce the cost of the instrument to make it truly portable and affordable for the patient.

FIGURE 4.2 Breezing metabolic tracker that can wirelessly transmit data through Bluetooth to personal smartphone for resting metabolic rate monitoring. Reproduced with permission of Breezing Co.

4.4.2 Blood Analyzer

Blood has plenty of biomarkers, which indicate certain diseases. Some portable devices have been developed to help patients to manage the diseases in home settings.

4.4.2.1 Glucose Meter Blood glucose monitoring is very important for diabetes treatment and management, especially for type 1 diabetes in which the pancreas of the patient produces very little or no insulin, insulin being the hormone that allows glucose to enter cells for energy production. The self-monitoring of glucose is recommended by the American Diabetes Association (ADA), the Food and Drugs Administration (FDA), and the National Institutes of Health (NIH) for patients with diabetes. Because of the prevalence of diabetes, the glucose test became the most frequently performed assay. Glucose biosensors contribute to about 85% of the entire biosensor market [23, 24]. Electrochemical techniques are the most commonly used technology for glucose detection, and the majority of the 6 billion annual assays performed by self-monitoring diabetic patient are electrochemical [23]. Traditionally, the glucose test is performed by piercing the skin to draw blood, then delivering certain amount of blood to a one-time-use "test-strip" of the glucose meter. This disposable "test-strip" based test does not have the continuous monitoring capability, and the finger piercing–sampling method could be uncomfortable to patients, which may limit patients for frequent monitoring. Developing glucose meters with the feature of continuous and noninvasive monitoring will be more suitable for the pervasive healthcare system because it will be less obstructive to patients and will provide dynamic information about the trend and pattern of the blood glucose level, including direction, magnitude, duration, and frequency, which could be very useful for medication and treatment. Continuous glucose monitor (CGM) is also essential for the closed-loop system design for diabetes treatment through which the real-time blood glucose regulating can be achieved by integrating with an insulin pump. The noninvasive glucose monitoring technologies based on optical and transdermal techniques have been proposed and are still under development [25]. Current CGMs on the market include the following: Guardian REAL-Time Continuous Glucose Monitoring System, Dexcom G4 Platinum, and MiniMed® 530G with Enlite, all of which are based on electrochemical techniques [26]. Additionally, a wireless transmitter is integrated in the unit for data transmission. This wireless data transmission capability makes them suitable to serve as an active sensor node in the BSN for pervasive health care. The sensor lifetime is 3–7 days. A common disadvantage of these CGMs is that they need frequent calibration (every few hours), which could be troublesome to the patient for daily use. There is also some concern regarding the accuracy that needs to be addressed in future R&D.

4.4.2.2 Blood Ketone Monitor Blood ketone monitoring is important for diagnosing diabetic ketoacidosis (DKA) in type 1 diabetes patients. The human body uses glucose as the main energy source. If the body is deficient of insulin, the blood glucose cannot enter the cells to produce energy. In this case, the body may break

down the other energy source, like fat, to generate energy to maintain body functions. Through this process, ketones are produced and begin to build up in the body (blood and urine) causing ketoacidosis. Monitoring blood ketone will enable the fast treatment of ketoacidosis by introducing extra insulin to the body, thus minimizing the health risk for the patient. Ketone monitoring can be achieved through either a urine or a blood test. The blood test, which measures β-hydroxybutyrate, is recommended by the ADA and has been reported to be more effective than the urine test [27]. Portable blood ketone meters based on electrochemical techniques have been developed for self-monitoring of ketones with a fast response time and high accuracy, including Nova Max plus and Precision Xtra. These blood ketone meters can help the patients with better diabetes management, reduce the patient morbidity, and save the healthcare cost [28]. Usually, these products integrated the blood glucose and blood ketone tests on the same sensor, which could potentially serve as multifunctional sensor node in the BSN for patient monitoring. However, they lack continuous monitoring capability and a wireless data transmission module. Adaptation and improvement need to be made on the current blood analyzers to make them more suitable to serve as the sensor nodes for the pervasive health monitoring system.

4.5 WEARABLE SENSORS

Wearable sensors are ideal candidates for the sensor node components because of their miniaturized size, fast response, continuous monitoring capability, wireless data transmission, and nonobstructive user experience. Moreover, they compromise the signal quality during monitoring and patient comfort. These features enable wearable sensors to long-term continuous physiological monitoring and could lead to better management and treatment of chronic illnesses, such as certain neurological, respiratory, and cardiovascular diseases. The common strategy in wearable sensors design is to integrate a tiny sensor into personal daily-use items, including but not limited to clothing, watches, glasses, shoes, wristbands, belts, and jewelry.

Usually wearable sensors refer to the sensing devices embedded in clothing or wearable accessories, such as watches, glasses, belts, shoes, jewelry, headgears, and arm-wears. According to "Wearable Technology 2014–2024: Technologies, Markets, Forecasts: ID TechEx," the wearable electronics business is projected to grow from over $ 14 billion in 2014 to over $ 70 billion in 2024 with health care becoming the dominant sector [29].

Since wearable sensors are seamlessly integrated to the daily-use wearable apparel, they are intrinsically nonobstructive and very suitable for continuous physiological parameters monitoring.

4.5.1 Wearable Sensing Technologies

Wearable sensors measure the physiological parameters generally through skin contact. Thus the sensing technologies adapted in wearable sensor design should be noninvasive and compatible to on-body sensing requirements. These sensing

technologies include, but not limited to, near-infrared (NIR) techniques, ECG techniques, electrochemical techniques, and accelerometry techniques.

1. Near-infrared techniques: NIR spectroscopy uses the electromagnetic spectrum in the NIR region (wavelength in the range of 800–2500 nm) to detect molecular overtone and combination vibrations. Since NIR can easily penetrate into biological tissues, it is suitable to detect physiological parameters in a noninvasive way. The NIR-based medical diagnostics include glucose concentration [30], oximetry [31], and even heart rate [32]. The glucose concentration in the tissues is measured by detecting the variations of light intensity due to transmission and reflectance, because the glucose concentration change can result in the absorption and scattering coefficients variation of the tissue. The blood oxygen saturation (SpO2) can be determined by measuring the NIR absorption characteristics of oxygenated and deoxygenated hemoglobin, the absorption spectra of which are 800–850 and 650–800 nm, respectively [33]. The continuous measurement of O_2 saturation in the tissue can be achieved by simultaneously monitoring NIR absorption of both oxyhemoglobin (HbO2) and deoxyhemoglobin. The heart rate can be determined from the HbO2 pulsation through NIR spectroscopy signal.
2. Electrocardiography techniques: ECG, which monitors the electrical activity of the heart, is a widely used vital sign sensing technology providing diagnostic information about the cardiovascular system. When the heart muscle depolarizes during the heartbeat, it will result in tiny electrical changes on the skin. The ECG monitor can detect the tiny voltage rise and falls caused by polarization and depolarization of the cardiac tissue and then generates the ECG waveforms. Typically multiple electrodes will be attached to specific parts of the chest, arms, hands, and legs for signal recording. In order to avoid the skin irritation and allergic reaction arising from the physical contact between the electrodes and the human skin, an on-contact ECG-sensing method has been developed by using small capacitive electrodes integrated in a cotton T-shirt [34]. By using wireless communication modules on the ECG platform, such as a radio-frequency transmitter [35], Bluetooth [36], or Zigbee [37], several wireless wearable ECG sensors have been developed, which can be used as sensor nodes for continuous arrhythmia monitoring in BSNs.
3. Electrochemical techniques: Electrochemical sensors detect the concentration of analytes by measuring the electrical signal generated during the electrochemical reaction of analytes on the electrodes surface. The elements of a typical electrochemical sensor consist of a working electrode, reference electrode, counter electrode, and electrolyte. Because of the miniaturized size, high sensitivity and selectivity, reliable performance, and low cost electrochemical sensors are ideal candidates for the sensor node in wireless BSNs. Electrochemical sensors have been developed to detect electrolytes and metabolites (including pH, fluoride, lactate, oxygen, glucose, ammonia, and sodium ion, ammonium ion, potassium ion, and chloride ion) in body fluids such as sweat and saliva in a real-time and noninvasive manner for health status monitoring [38]. In order to make the

electrochemical sensor wearable, flexible substrates like fabric and plastic have been applied to electrochemical sensor design and fabrication [39].

4. Accelerometry techniques: Accelerometers with single and multiaxis modes can detect the magnitude and direction of the acceleration. It can be used to detect orientation, coordinate acceleration, and vibration. Accelerometers have wide applications. One of the new applications is to measure human acceleration for clinical assessments. The most popular application of accelerometer is the monitoring of physical activity and assessment of metabolic energy expenditure [40]. By locating the accelerometer to different locations of human body, some other health information can also be obtained. When placing the accelerometer to the wrist, the Parkinsonian bradykinesia can be monitored, while the coughing could be detected when the accelerometer is attached to the chest. Another promising application is to use an accelerometer to detect the falls of elderly people [41], which could potentially prevent the falls and reduce the response time of paramedics for treatment. Other applications of accelerometers include balance and postural sway, gait, and movement classification.

4.5.2 Wearable Sensors for Chronic Disease Management

Wearable sensing technologies have wide applications in the chronic disease management, such as diabetes diagnosis, cardiovascular disease diagnosis, and neurological disease diagnosis.

4.5.2.1 Wearable Sensors for Diabetes Diagnosis Blood glucose monitoring is the key step for diabetes diagnosis. Products with the capability of continuous glucose monitoring that are wearable will be revolutionary to the market. Abbott's Freestyle Libre Flash Glucose Monitoring System is a wearable glucose monitor that can be worn on the back of the upper arm for glucose monitoring for up to 14 days [42]. There is no finger prick and calibration needed, which makes it unique compared with current glucose monitoring systems. The Abbott's Freestyle Libre System uses a sensor, which is of the size of a two Euro coin, for glucose measurement in every minute. This small sensing unit has a small filament inserted under the skin to get interstitial fluid, and the wearable feature is ensured by using a small adhesive pad for fixing the position. The system has a reader to get the glucose readings, display the real-time result, and record the history data. Up to 90 days of data can be saved, and the glucose profile of the patient can be shared between patient and healthcare professionals.

The noninvasive feature is highly preferred for wearable glucose monitor. Optical technique can be used for noninvasive glucose detection by detecting the interaction between light and glucose molecules. Infrared absorption spectroscopy has been studied for the quantitative determination of glucose concentration in blood [43]. GlucoWise™, a noninvasive and wireless glucose monitor is able to extract blood glucose levels every few seconds by transmitting low-power radio waves through a finger or an earlobe [44]. The glucose information can be wirelessly uploaded to the

mobile app and the cloud for history tracking. GlucoWise is currently in development and will be available on the market once clinical trials are completed.

Google announced in 2004 that they are developing a "smart" contact lens to help diabetics to monitor their glucose levels [45]. The lens is designed to measure glucose in tears continuously using a wireless chip in a less invasive way. The RF identification (RFID) wireless chip and miniaturized glucose sensors are embedded between two layers of soft contact lens material, and a pinhole in the lens is designed to allow fluid from the surface of the eye to permeate into the sensor. According to the design, the power and data transmission comes wirelessly from the RFID capability. The integrated electronics in the lens will not obscure vision since they locate outside the eye's pupil and iris.

4.5.2.2 Wearable Sensor for Cardiovascular Disease Diagnosis Cardiovascular disease accounted for 31.9% of all deaths in the United States in 2010 [46]. Continuous monitoring of a patient can help doctors to track disease progression and provide earlier treatment. For effective cardiovascular disease diagnosis, physiological parameters such as ECG, blood pressure, and heart rate need to be monitored.

It has been reported that ECG can be used to predict the risk of death in the patients with cardiovascular disease [47]. Many wearable ECG sensors have been developed for continuous and wireless monitoring of patients [48, 49]. Rune Fensli et al. have developed a wearable ECG-recording sensor for continuous arrhythmia monitoring [35]. The system includes two electrical contact points for sensing, a built-in RF-radio transmitter for wireless data communication, an arrhythmia algorithm based on nonlinear transformation as alarm criteria, and a clinical application to allow the doctor to overview the latest alarm recording and make comments. Byungkook Jeon et al. designed and implemented a smart shirt with an ECG sensor that can be worn by patients for real-time monitoring [50]. This wearable ECG system works with an app based on the Android OS platform for real-time monitoring, self-diagnosis, and remote diagnosis for chronic heart disease patients. Healthcare professionals can access patient's data wirelessly in real time with their smartphones. This proposed system could be implemented in a pervasive healthcare environment for chronic heart disease patients. A wearable and wireless ECG monitor, QardioCore (Fig. 4.3), which can improve the detection of cardiac conditions in daily life, is expected to be available on the market soon [51].

Traditional blood pressure measurements can only provide "snap-shot" information, while wearable sensors, in the form of a wristwatch or jewelry, can provide long-term continuous blood pressure monitoring, which can improve the diagnosis and treatment of cardiovascular diseases such as hypertension, heart failure, and other cardiovascular disorders. A wearable and unobtrusive blood pressure sensor based on a ferroelectric film for long-term continuous monitoring of patients has been developed by researchers from Korea [52]. The company, STBL Medical Research AG, has designed and implemented a wristwatch-like device for continuous blood pressure recording. Several sensors were used to simultaneously measure the contact pressure, pulse, and blood flow on the skin of the wrist. Since this device is wearable and conformable to the patients, it can record blood pressure in the patient's natural

FIGURE 4.3 QardioCore wearable EKG/ECG monitor. The data can be automatically shared with doctor and family. Reproduced with permission of Qardio, Inc.

environment and reduce the measurement distortion due to the nervousness of the patients [53].

Lots of wearable heart rate monitors have been developed and are available on the market in the form of watches, including Suunto Quest, Polar RS300X, and Garmin Forerunner 15 [54], and headphones for real-time continuous heart rate monitoring. In these products, the heart rate signal can be picked up by a transmitter in the sensor based on the electrical signal transmitted through the heart muscle during heartbeat. The signal can be captured either from the chest through a chest strap or from the wrist through a watch. Some products can wirelessly communicate with the phone app through Bluetooth for data transmission. Intel's SMS Audio Biosport In-Ear Headphones can monitor the heart rate consistently through a unique in-ear heart rate monitor. This self-charging system can also sync with the fitness app Runkeeper for the heart rate, pace distance, elevation, and calorie burn [55].

4.5.2.3 Wearable Sensors for Respiratory Disease Diagnosis As for respiratory disease screening, patient monitoring, and disease management, pulse oximetry is a useful and convenient tool since it is noninvasive and suitable for continuous monitoring [56]. Patients with pulmonary diseases such as COPD and sleep apnea syndrome may have low oxygen saturation, which can be measured by an oximeter. Studies have indicated that pulse oximetry is helpful in screening of the patients with COPD and identifying the patients who might benefit from long-term oxygen therapy [57]. Pulse oximetry can also be used to assess the severity of asthma attacks and the effectiveness of the treatment for asthma patients by complementing the peak flow

meters [58]. Pulse oximetry is also useful for the assessment and management of respiratory rehabilitation and therapy.

Fingertip-type pulse oximeters are the most popular pulse oximeters in the current market [59], but are not appropriate for long-term monitoring since they are obstructive to the users and interferes with their daily activity. These products measure SpO_2 continuously at the fingertip and connect to a wrist band for signal processing and result display. They use both red and infrared light to detect oxygenated and deoxygenated hemoglobin, based on which the oxygen in the blood is determined. Some oximeters have the Bluetooth communication capability to sync with a health app for data management, such as the iHealth Wrist-Worn Pulse Oximeter [60].

Ring-type wireless oximeters are less obstructive and more convenient for long-term continuous monitoring. Yang et al. developed a miniaturized oximeter in a ring configuration [61]. This ring sensor can be worn by the patient for monitoring health status 24 h a day. It uses optoelectrical components for oxygen saturation monitoring in a finger base and a wireless transmitter for data communication between the sensor and the host computer. Huang et al. also developed an optimal design of a ring-type pulse oximeter using multiple detectors to improve the efficiency of the blood oxygen measurement [59].

Besides ring-type oximeters, a wearable electronic patch with reflectance pulse oximetry has also been developed by Rasmus G. Haahr et al. Based on this device physiological information could be provided for heart and lung function assessment [62].

4.5.2.4 Wearable Sensors for Neurological Disease Diagnosis Parkinson's disease is a neurodegenerative disorder of the central nervous system causing multiple impairments, such as bradykinesia (i.e., slowness of movement), tremor, rigidity (i.e., resistance to externally imposed movements), flexed posture, postural instability, and freezing of gait. Parkinson's disease is more commonly seen in elderly people, and it currently affects about 2% of the population over the age of 65, which is expected to double by 2020 [63]. Wearable sensors are suitable for remote and continuous health status monitoring of patients with Parkinson's disease and can provide accurate information about the disease evolution based on data processing, which offers the qualitative and quantitative assessment and treatment for personal disease management.

Accelerometer-based wearable sensors for Parkinson's disease monitoring have been intensively developed and studied, including wrist- or belt-mounted activity monitors [64, 65], ankle-mounted sensor arrays [66], and on-shoe wearable sensors [67]. In an ideal wearable sensing platform, the following functions should be provided: (i) carrying out effective evaluation of the symptoms via unobtrusive and continuous monitoring, (ii) building up the personal disease profile of the patient, and (iii) involving healthcare professionals in the system to provide effective and timely therapy suggestions to the patients.

Furthermore, wireless accelerometer nodes and smartphones, such as the iPhone, can also be used for optimizing therapy strategy by providing convenient quantified feedback [68]. Filippo Casamassima et al. developed a system based on a wireless

BSN and a smartphone for gait training and rehabilitation for Parkinson's disease patients in daily life [69]. This system has adopted a closed-loop approach to enhance motor learning in specific tasks for patients with Parkinson's disease. The system enables real-time extraction of gait spatiotemporal features and their comparison with a patient's reference walking parameters and returns the feedback to the user in the form of vocal messages, encouraging the user to keep their walking behavior or to correct it.

Besides accelerometer-based wearable sensors for movement monitoring in patients with neurological disease, electrodermal activity (EDA) sensors could also be used for unobtrusive wearable device development. Ming-ZherPoh reported a wearable wristband integrated with an EDA sensor that provides comfortable and continuous EDA measurements during daily activities. This system can be potentially used for diagnosis and understanding of neurological conditions [70].

Several products have also been introduced to the market to benefit patients with Parkinson's disease. Personal KinetiGraph™, a wrist-worn medical device developed by Global Kinetics Corporation, can automatically collect the movement data of the patients with Parkinson's disease, remind the patient to take medication, and help doctors to identify changes in the movement symptoms for therapy optimization. The data file of the patient can be uploaded from this wrist worn device to the tablet and finally sent to the server for processing. A patient report will be generated, sent back to the patient, and emailed to the clinician for an assessment. This product received FDA clearance already [71]. Intel announced a partnership with the Michael J. Fox Foundation for Parkinson's Research to combine their wearable device with analytics software for better recording a patient's symptoms and offering new insight into the disease and treatment. The big data analytics will help to gather real-time patient data from wearable devices for aggregation and analysis in order to better quantify diagnosis and treatment [72].

4.6 IMPLANTABLE SENSORS

Wearable sensors are attractive because of their noninvasive feature, but in some cases implantable sensors are preferred. Though the patient needs to undergo an invasive procedure, the implantable sensors can provide the *in vivo* diagnostic without any intervention from the patients because the sensors have direct interface with the internal organs or tissues. The technological advance in miniaturization of sensing units and energy source makes the implantable sensor highly practical for medical applications. Once implanted in the human body, the sensor node will continuously monitor the physiological parameter and wirelessly transmit the data to the remote center. Based on the real-time data, decisions will be made to guide the patient to an effective treatment.

Most of the implantable sensors are focusing on the cardiac monitoring for the patients who suffer from arrhythmias and require long-term continuous monitoring. The Reveal LINQ™ insertable cardiac monitoring system from Medtronic is an insertable cardiac monitor designed to help the doctor diagnose and treat irregular

heartbeats that may be related to unexplained fainting [73]. It is implanted just under the skin of the patient's chest and can automatically detect and record abnormal heart rhythms for up to 3 years.

Implantable glucose sensors such as GlySens ICGM™ System have also been brought to the market for long-term continuous glucose monitoring to improve the lives of people with diabetes [74]. The GlySens ICGM System is based on glucose-specific, oxygen-based, dual-enzyme electrode technology, which is comprised of two units: (i) a fully implanted sensor (designed for long operational life, up to 18 months) and (ii) an external receiver with monitor. The external receiver is designed to provide a convenient way for continuous, at-a-glance glucose display, recording, and alerts regarding hypo- or hyperglycemic glucose excursions. It is noteworthy that the fully implanted GlySens sensor wirelessly communicates with the external receiver, which offers high-bandwidth data source potentially applicable to new applications of diabetes-related health information technology. This feature provides the possibility of serving as the sensor node in the BSN to wirelessly transmit a real-time glucose profile of the patient.

Besides the products available on the market, ongoing researches have been carried out to develop new implantable sensors, and future products are foreseeable in the pipeline. Implantable chemical sensors for real-time monitoring of clinically important species such as oxygen partial pressure (PO2), carbon dioxide partial pressure (PCO2), pH, lactate, sodium ion, potassium ion, calcium ion, and oxygen saturation by applying electrochemical and optical techniques have been reported [75]. The barrier between the early proof-of-concept prototype and the final product is the reliability issue of the analytical devices after implantation.

4.7 WIRELESS COMMUNICATION

Two kinds of data transmissions are involved in the pervasive computing architecture for healthcare service. First, the physiological data collected from individual sensor node must be transmitted to the MBU for processing and organizing. Second, the data accumulated in the MBU need to be sent out through the gateway to the health-care professionals or cloud.

The data communication between the sensor node and the MBU gateway is typically within the house range. Transmission technologies like Bluetooth, Zigbee, and RFID are often used.

1. Bluetooth: Bluetooth technology is widely used in the current sensor devices for the following reasons: (i) The data transmission rate is very high, which means more information can be transferred and fast transmission can be achieved (ms), and in most cases real-time data transmission can be achieved. (ii) The data transmission is very reliable and the chance of dropping bytes is low, especially when a robust communication protocol is defined between the sensor node and the MBU. (iii) The signal is strong enough to cover the house-range data communication and a communication range of 13 m can be

guaranteed with the newest technology. (iv) The Bluetooth chip is very small (about the size of a coin) that can make the sensor node small, which increases the portability and comfort for wearing. (v) The Bluetooth technology is relatively cheap compared to other transmission solutions (a Bluetooth chip is just a few dollars). (vi) It has the highest compatibility with the current smartphone design, and almost all smartphones have built-in Bluetooth modules. Thus the smartphone can be the most common and pervasive MBU to communicate with the sensor node array to gather the health data and the gateway to push the health information into the cloud and server. However, the power consumption of Bluetooth is relatively high, in the range of 50–100 mW. According to the master-slave protocol, the master can communicate to maximum seven slaves, which may not be enough to support larger number of sensor nodes in the BSN system. Bluetooth has been widely used in lots of wearable sensor prototypes, such as GlucoWise glucose monitor [44] and ECG monitors [36]. Bluetooth-based sensor networks have been reported for remotely monitoring physiological signals of a patient [76].

2. Zigbee: Zigbee is a communication specification based on IEEE 802.15.4 protocol. Zigbee is suitable for mobile health applications in wireless BSN because of the following features: low power consumption, long transmission distance (<1 mW for 30 m range), high security (128-bit encryption keys), capability of communicating with a large number of sensor nodes (support thousands of devices), support for rapid disconnect and rejoin, and cost effective (compared to Bluetooth or Wi-Fi for the wireless BSN). Zigbee-based ECG monitoring systems [37] and ring-type pulse oximeters [77] have been developed for healthcare application. Ubiquitous healthcare services using Zigbee and mobile phones for elderly patients with diabetes mellitus or heart disease have been reported. Zigbee could be one of the potential candidates for wireless healthcare systems due to its lower power consumption [78].
3. RFID: RFID techniques can also be used for wireless data transmission in the wireless BSN. Based on RFID technology, a wireless and battery-free sensor node can be created for sensing and computation, for example, temperature sensors and accelerometers [79]. A cost-effective general-purpose multi-ID tag has been developed to wirelessly transmit whatever kind of data in sensor networks [80]. Google is developing a “smart” contact lens to help people with diabetes monitor their glucose levels by using RFID for wireless power and data transmission.

The data communication between the personal gateway and the cloud or server in the Internet or database can be achieved only through the Internet or mobile phone network, which highly depends on the infrastructure and the service providers.

The health data needs to be securely stored on a server or platform, from which authorized health professionals should analyze the data and provide diagnosis and timely medical feedback and assistance to the patients. In the ideal situation, the prescription and dosage of the medication can be immediately sent to the patient so that the treatment will be carried out in a timely and efficient manner.

4.8 MOBILE BASE UNIT

The MBU is also called LPU. As mentioned before, the MBU serves as the gateway to bridge the personal wireless sensor networks and the healthcare providers. Unlike wireless sensor networks that focus on the physiological parameter monitoring and data collecting, the MBU takes responsibility for data aggregation, data handling, and communication. It also creates a user interface to help the healthcare professionals to analyze the monitoring parameters according to the patient's personal profile.

MBU can be implemented on PDAs or mobile phones by running software applications. The PDA and mobile phones are ideal candidates for the MBU because of the following reasons: (i) they have built-in wireless communication modules, such as Bluetooth, which can easily communicate with the sensor nodes in the BSN to collect physiological parameters; (ii) they can serve as a gateway to communicate with the healthcare server to share health data and get treatment suggestions through the Internet or 4G; (iii) they have powerful computing capabilities for health data processing, aggregation, and storage; (iv) they are available to most of the users and will not cause extra financial burden to them; (v) software apps can be easily developed, maintained, and updated to provide adaptive services; and (vi) they are portable.

The data processing through the MBU can be achieved in two ways: on-board processing and on-server data processing. On-board data processing uses the computation capability of a local MBU to process the data and provide feedback to the user according to certain algorithm design, while the on-server processing needs to send the local data to a remote server for processing and generating feedback. The local data processing can handle data in lower order and provide immediate alarm service to the patient in case of an emergency. Since it is independent of the network operation, the feedback to the patients can be timely and efficient. The remote data processing on the server allows big data computation, which is suitable for dealing with complex conditions and for providing an overall assessment of the patient [81, 82].

It should be mentioned that the data security is also a very important topic in the BSN–MBU system. The health data of the patient must be kept as confidential information to ensure the respect for privacy of the patient. Only the authorized personnel are allowed to access the health database of the patient. The integrity and authenticity of the health data should also be guaranteed in order to provide the better diagnosis and treatment. As what Ramon Martí et al. described the requirements and implementation of security mechanism for a wireless mobile healthcare system in their paper, the security requirements include the following: (i) security in all the hardware devices in the system, (ii) security in data transmission, and (iii) security in data storage [83].

Marcin Bajorek et al. proposed a home-monitoring healthcare system for continuous monitoring of biomedical signals with an Android mobile tablet serving as the MBU for data exchange and visualization. The software of this system was developed in the Java/Jaca FX environment, and security mechanisms such as device

registration, data encryption, and special system structure design were applied to ensure system security [84]. V.M. Jones et al. also designed an epilepsy monitoring system using an HTC P3600 smartphone as the MBU. This system can process the biosignals and context information to identify medical emergencies and facilitate an appropriate response [85]. Other health monitoring systems with the PDA/cell phone as the MBU also have been proposed [86–89].

4.9 CONCLUSION AND CHALLENGES

The technology advancements of PDAs/smartphones, portable sensors, wearable sensors, and implantable sensors make the constant monitoring of patients practical and attractive for those suffering from chronic diseases, such as certain neurological, respiratory, and cardiovascular diseases. Intensive studies on the monitoring system design, new sensing platform development, new application scenarios exploration, data processing algorithms design, and security mechanisms implementation have been carried out, and tremendous progress has been made in this field. There is no doubt that the diagnosis and treatment service provided by the pervasive healthcare system based on wireless BSNs will dramatically improve the quality of life of the patients and significantly reduce the healthcare costs on the patients and society [90].

However, there are still challenges and hurdles for turning the pervasive health care from concept to real application to benefit the patients and the entire society. Up to now, most of the researches in this area are still limited to the laboratory settings or artificial clinical environments. Though some of the diagnostic products are available on the market, they can only provide a certain type of physiological information, rather than a complete physiological profile of the patient over a long period of time, and in most cases there is a lack of feedback services from healthcare professionals. Some of the individual sensors are still bulky in size and require patient's interaction for proper use, which may need further improvement to serve as the sensor node of the BSN. The MBU/BSN model relies on wearable and wireless sensor nodes for health information collection and transmission, thus the device safety and user privacy could be important concerns. Furthermore, the standardization and infrastructure construction are also critical for the implementation of the MBU/BSN model in the healthcare system.

To bring the pervasive healthcare system from research to the practical application, major challenges include, but are not limited to, the following:

1. Compromise between comfort and accuracy: To achieve the goal of continuous monitoring, the sensor nodes must be fully miniaturized and embedded into wearable sensors so that they can be nonintrusive to the patients and not affect the patients' daily life. The accuracy of the sensor node could be an issue when the sensors are worn by the patients during daily activities. The mechanical movements of the patients and the instability of the contact between sensors and human body could introduce noise and interference to the sensor signal. The technical advance in materials (e-textiles and flexible and stretchable

electronics) and smart wearable product design (head-worn, straps, shirts, wrist-worn, clips, shoe-worn/foot pods) could be helpful to solve this problem. But solutions from the signal processing, algorithm design, and ergonomics are also needed to address this issue.

2. Power for the BSN: BSNs use multiple sensor nodes for continuous monitoring of patients and wireless communication with an MBU. Thus, the total power consumption could be significant. If batteries are used for power supply, the size and weight of the batteries can be a burden to the patient. And also the patient needs to recharge the batteries for long-term use, which is not user-friendly. Energy harvesting and storage technologies could be a cost-effective and environment-friendly solution to power the sensor nodes in the BSN. But the technical barriers and limitations need to be overcome to make these technologies reliable and practical.
3. Health data interpretation: A larger amount of continuous health data will be collected through the BSN from the patient. How to efficiently analyze the data and correlate different physiological data to create a comprehensive health profile for the patient? Considering the diversity of the patients, how to create a reliable alert mechanism? How to generate personalized treatment and therapy plans? These questions need to be answered to provide high-quality medical service to the patients.

More efforts are desired from both academic and industrial fields to address the challenges already mentioned so that the pervasive health care can be realized in the current medical care system to provide more efficient, cost-effective diagnosis, and treatment services to patients.

ACKNOWLEDGMENT

Thanks go to my colleagues Nongjian Tao, Erica Forzani, Francis Tsow, and Devon Bridgeman for help and discussion.

REFERENCES

[1] Arnrich B., Mayora O., Bardram J., Tröster G., Pervasive healthcare: paving the way for a pervasive, user-centered and preventive healthcare model. Methods Inf Med 49, 67–73 (2010).

[2] Varshney U., Pervasive healthcare and wireless health monitoring. Mobile Netw Appl 12, 113–127 (2007).

[3] Den Hond E., Celis H., Vandenhoven G., O'Brien E., Staessen J. A., Determinants of white-coat syndrome assessed by ambulatory blood pressure or self-measured home blood pressure. Blood Press Monit 8, 37–40 (2003).

[4] Heller A., Integrated medical feedback systems for drug delivery. AIChE J 51, 1054–1066 (2005).

[5] Zhang M., Sawchuk A. A., A Customizable Framework of Body Area Sensor Network for Rehabilitation. Proceedings of the 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), Bratislava, November 24–27, (2009), IEEE, pp. 1–6.

[6] Pawar P., Jones V., van Beijnum B. J., Hermens H., A framework for the comparison of mobile patient monitoring systems. J Biomed Inform 45, 544–556 (2012).

[7] Triantafyllidis A., Koutkias V., Chouvarda I., Maglaveras N., An open and reconfigurable wireless sensor network for pervasive health monitoring. Methods Inf Med 47, 229–234 (2008).

[8] Lecointre A., Dragomirescu D., Dubuc D., Katia G., Patrick P., Aubert H., Muller A., Berthou P., Gayraud T., Plana R., Miniaturized Wireless Sensor Network. eprint arXiv:1002.0507 (2010).

[9] Hum A. P., Fabric area network: a new wireless communications infrastructure to enable ubiquitous networking and sensing on intelligent clothing. Comput Netw 35, 391–399 (2001).

[10] Alving K., Weitzberg E., Lundberg J. M., Increased amount of nitric oxide in exhaled air of asthmatics. Eur Respir J 6, 1368–1370 (1993).

[11] Rutgers S. R., Postma D. S., van der Mark T. W., Koëter G. H., Nitric oxide in exhaled air in COPD. Eur Respir J 9, 13s (1996).

[12] Dötsch J., Demirakça S., Terbrack H. G., Hüls G., Rascher W., Kühl P. G., Airway nitric oxide in asthmatic children and patients with cystic fibrosis. Eur Respir J 9, 2537–2540 (1996).

[13] Kharitonov S., Alving K., Barnes P. J., Exhaled and nasal nitric oxide measurements: recommendations. Eur Respir J 10, 1683–1693 (1997).

[14] Hemmingsson T., Linnarsson D., Gambert R., Novel hand-held device for exhaled nitric oxide-analysis in research and clinical applications. J Clin Monit Comput 18, 379–387 (2004).

[15] Haugen H. A., Chan L. N., Li F., Indirect calorimetry: a practical guide for clinicians. Nutr Clin Pract 22, 377–388 (2007).

[16] Harris J. A., Benedict F. G., A biometric study of human basal metabolism. Proc Natl Acad Sci U S A 4, 370–373 (1918).

[17] Frankenfield D., Roth-Yousey L., Compher C., Comparison of predictive equations for resting metabolic rate in healthy nonobese and obese adults: a systematic review. J Am Diet Assoc 105, 775–789 (2005).

[18] Weir J. B., New methods for calculating metabolic rate with special reference to protein metabolism. J Physiol 109, 1–9 (1949).

[19] D. Wilcoxson, Microlife BodyGem and MedGem Indirect Calorimeters. http://metabolicratetest.com/. Accessed January 2, 2016.

[20] Breezing, Measure your metabolism using the world's only mobile metabolism tracker. http://breezing.com/. Accessed January 2, 2016.

[21] Phillips M., Altorki N., Austin J. H., Cameron R. B., Cataneo R. N., Greenberg J., Kloss R., Maxfield R. A., Munawar M. I., Pass H. I., Rashid A., Rom W. N., Schmitt P., Prediction of lung cancer using volatile biomarkers in breath. Cancer Biomark 3, 95–109 (2007).

[22] Phillips M., Cataneo R. N., Ditkoff B. A., Fisher P., Greenberg J., Gunawardena R., Kwon C. S., Tietje O., Wong C., Prediction of breast cancer using volatile biomarkers in the breath. Breast Cancer Res Treat 99, 19–21 (2006).

[23] Heller A., Feldman B., Electrochemical glucose sensors and their applications in diabetes management. Chem Rev 108, 2482–2505 (2008).

[24] Wang J., Electrochemical glucose biosensors. Chem Rev 108, 814–825 (2008).

[25] Oliver N. S., Toumazou C., Cass A. E., Johnston D. G., Glucose sensors: a review of current and emerging technology. Diabet Med 26, 197–210 (2009).

[26] Comparison of Current Continuous Glucose Monitors (CGMs), 2014. http://www.diabetesnet.com/diabetes-technology/meters-monitors/continuous-monitors/compare-current-monitors. Accessed January 4, 2016.

[27] Klocker A. A., Phelan H., Twigg S. M., Craig M. E., Blood β-hydroxybutyrate vs. urine acetoacetate testing for the prevention and management of ketoacidosis in Type 1 diabetes: a systematic review. Diabet Med 30, 818–824 (2013).

[28] Marlyn L., Role of blood ketone testing in sick-day management. Manag Care 13, 11–14 (2004).

[29] P. Harrop, J. Hayward, R. Das, and G. Holland, Wearable technology 2015–2025: technologies, markets, forecasts, 2015. http://www.idtechex.com/research/reports/wearable-technology-2014-2024-technologies-markets-forecasts-000379.asp. Accessed January 4, 2016.

[30] Vashist S. K., Non-invasive glucose monitoring technology in diabetes management: a review. Anal Chim Acta 750, 16–27 (2012).

[31] Ferrari M., Quaresima V., Review: near infrared brain and muscle oximetry: from the discovery to current applications. J Near Infrared Spectrosc20, 1–14 (2012).

[32] Perdue K. L., Westerlund A., McCormick S. A., Nelson C. A., Extraction of heart rate from functional near-infrared spectroscopy in infants. J Biomed Opt 19, 067010 (2014).

[33] Murkin J. M., Near infrared cerebral oxygenation monitoring. Appl Cardiopulm Pathophysiol 13, 152–154 (2009).

[34] Nemati E., Deen M.J., Mondal T., A wireless wearable ECG sensor for long-term applications. IEEE Commun Mag 50, 36–43 (2012).

[35] Fensli R., Gunnarson E., Gundersen T., A Wearable ECG-Recording System for Continuous Arrhythmia Monitoring in a Wireless Tele-Home-Care Situation. Proceedings of 18th IEEE Symposium on Computer-Based Medical Systems, Dublin, June 23–24, (2005), IEEE, pp. 407–412. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1467724&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1467724. Accessed February 19, 2016.

[36] Gargiulo G., Bifulco P., Cesarelli M., Ruffo M., Romano M., Calvo R. A., Jin C., van Schaik A., An ultra-high input impedance ECG amplifier for long-term monitoring of athletes. Med Devices 3, 1–9 (2010).

[37] Lee S. H., Jung S. M., Lee C. K., Jeong K. S., Cho G., Yoo S. K., Wearable ECG Monitoring System Using Conductive Fabrics and Active Electrodes. in Institute of Health Informatics, University of Minnesota, Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction, Lecture Notes in Computer Science 5612, Springer, Heidelberg, 778–783 (2009). https://www.researchgate.net/publication/221096580_Wearable_ECG_Monitoring_System_Using_Conductive_Fabrics_and_Active_Electrodes. Accessed February 19, 2016.

[38] Bandodkar A. J., Wang J., Non-invasive wearable electrochemical sensors: a review. Trends Biotechnol 32, 363–371 (2014).

[39] Jia W., Bandodkar A. J., Valdés-Ramírez G., Windmiller J. R., Yang Z., Ramírez J., Chan G., Wang J., Electrochemical tattoo biosensors for real-time noninvasive lactate monitoring in human perspiration. Anal Chem 85, 6553–6560 (2013).

[40] Mathie M. J., Coster A. C. F., Lovell N. H., Celler B. G., Detection of daily physical activities using a triaxial accelerometer. Med Biol Eng Comput 41, 296–301 (2003).

[41] Doughty K., Lewis R., McIntosh A., The design of a practical and reliable fall detector for community and institutional telecare. J Telemed Telecare 6 Suppl 1, S150–S154 (2000).

[42] Abbott. Mark for FreeStyle® Libre, a Revolutionary Glucose Monitoring System for People with Diabetes, 2014. https://www.abbottdiabetescare.com/press-room/2014/2014-e.html. Accessed January 4, 2016.

[43] Shen Y. C., Davies A. G., Linfleld E. H., Elsey T. S., Taday P. F., Arnone D. D., The use of Fourier-transform infrared spectroscopy for the quantitative determination of glucose concentration in whole blood. Phys Med Biol 48, 2023–2032 (2003).

[44] GlusoWise. Imagine living a healthier life with GlucoWise™. http://www.gluco-wise.com/. Accessed January 4, 2016.

[45] foxnews.com. Google announces 'smart' contact lenses that monitor glucose levels, 2014. http://www.foxnews.com/tech/2014/01/16/google-announces-contact-lens-glucose-monitor/. Accessed January 4, 2016.

[46] Go A. S., Mozaffarian D., Roger V. L., Benjamin E. J., Berry J. D., Blaha M. J., Dai S., Dai S., Ford E. S., Fox C. S., Franco S., Fullerton H. J., Gillespie C., Hailpern S. M., Heit J. A., Howard V. J., Huffman M. D., Judd S. E., Kissela B. M., Kittner S. J., Lackland D. T., Lichtman J. H., Lisabeth L. D., Mackey R. H., Magid D. J., Marcus G. M., Marelli A., Matchar D. B., McGuire D. K., Mohler III E. R., Moy C. S., Mussolino M. E., Neumar R. W., Nichol G., Pandey D. K., Paynter N. P., Reeves M. J., Sorlie P. D., Stein J., Towfighi A., Turan T. N., Virani S. S., Wong N. D., Woo D., and Turner M. B., Heart disease and stroke statistics: 2014 update. A report from the American Heart Association. Circulation 129, E28–E292 (2014).

[47] Syed Z., Scirica B. M., Mohanavelu S., Sung P., Michelson E. L., Cannon C. P., Stone P. H., Stultz C. M., Guttag J. V., Relation of death within 90 days of non-ST-elevation acute coronary syndromes to variability in electrocardiographic morphology. Am J Cardiol 103, 307–311 (2009).

[48] Lobodzinski S. S., Laks M. M., New devices for very long-term ECG monitoring. Cardiol J 19, 210–214 (2012).

[49] Deopujari S. P., Gaikwad A., Light weight, low cost, wearable ECG monitoring. Int J Comput Eng Res 3, 1–6 (2013).

[50] Jeon B., Lee J., Choi J., Design and implementation of a wearable ECG system. Int J Smart Home 7, 61–70 (2013).

[51] Qardio, QuardioCore, 2016. http://www.getqardio.com/qardiocore-wearable-ecg-ekg-monitor-iphone/. Accessed January 4, 2016.

[52] Noh S., Yoon C., Hyun E., Yoon H. N., Chung T. J., Park K. S., Kim H. C., Ferroelectret film-based patchtype sensor for continuous blood pressure monitoring. Electron Lett 50, 143–144 (2014).

[53] CORNELIA ZOGG, Wristband revolutionises blood pressure measurement, 2013. http://www.empa.ch/plugin/template/empa/3/137220/---/l=2. Accessed January 4, 2016.

[54] Gordon B.. Heart rate monitor watches review, 2016. http://heart-rate-monitor-watches-review.toptenreviews.com/. Accessed January 4, 2016.

[55] Intel, Smart earbuds with an integrated heart monitor. http://www.intel.com/content/www/us/en/wearables/fitness.html. Accessed January 4, 2016.

[56] Chavannes N., Pulse oximetry and respiratory disease in primary care. Prim Care Respir J 12, 2–3 (2003).

[57] Roberts C. M., Franklin J., O'Neill A., Roberts R. P., Ide J., Hanley M. L., Edwards J., Screening patients in general practice with COPD for long-term domicillary oxygen requirement using pulse oximetry. Respir Med 92, 1265–1268 (1998).

[58] Geelhoed G. C., Landau L. I., LeSouëf P. N., Oximetry and peak expiratory flow in assessment of acute childhood asthma. J Pediatr 117, 907–909 (1990).

[59] Huang C. Y., Chan M. C., Chen C. Y., Lin B. S., Novel wearable and wireless ring-type pulse oximeter with multi-detectors. Sensors 14, 17586–17599 (2014).

[60] Gaurav, iHealth Launches New Wristworn Pulse Oximeter, Ambulatory Heart and Blood Pressure Monitors at CES 2014, 2014. http://www.medgadget.com/2014/01/ihealth-launches-new-wristworn-pulse-oximeter-ambulatory-heart-and-blood-pressure-monitors-at-ces-2014.html. Accessed January 4, 2016.

[61] Yang B. H., Rhee S., Development of the ring sensor for healthcare automation. Robot Auton Syst 30, 273–281 (2000).

[62] Haahr R. G., Duun S. B., Toft M. H., Belhage B., Larsen J., Birkelund K., Thomsen E. V., An electronic patch for wearable health monitoring by reflectance pulse oximetry. IEEE Trans Biomed Circuits Syst 6, 1–9 (2012).

[63] Lau L. M., Breteler M., Epidemiology of Parkinson's disease. Lancet Neurol 5, 525–535 (2006).

[64] Saito N., Yamamoto T., Sugiura Y., Shimizu S., Shimizu M., Lifecorder: a new device for the long-term monitoring of motor activities for Parkinson's disease. Intern Med 43, 685–692 (2004).

[65] Hoff J. I., van der Meer V., van Hilten J. J., Accuracy of objective ambulatory accelerometry in detecting motor complications in patients with Parkinson's disease. Clin Neuropharmacol 27, 53–57 (2004).

[66] Moore S. T., MacDougall H. G., Gracies J. M., Cohen H. S., Ondo W. G., Long-term monitoring of gait in Parkinson's disease. Gait Posture 26, 200–207 (2007).

[67] Mariani B., Jiménez M. C., Vingerhoets F. J., Aminian K., On-shoe wearable sensors for gait and turning assessment of patients with Parkinson's disease. IEEE Trans Biomed Eng 60, 155–158 (2013).

[68] LeMoyne R., Wearable and wireless accelerometer systems for monitoring Parkinson's disease patients: a perspective review. Adv Parkinson's Dis 2, 113–115 (2013).

[69] Casamassima F., Ferrari A., Milosevic B., Ginis P., Farella E., Rocchi L., A wearable system for gait training in subjects with Parkinson's disease. Sensors 14, 6229–6246 (2014).

[70] Poh M., Swenson N. C., Picard R. W., A wearable sensor for unobtrusive, long-term, assessment of electrodermal activity. IEEE Trans Biomed Eng 57, 1243–1252 (2010).

[71] Global Kinetics Corporation. Parkinson's KinetiGraph™, 2013. http://www.globalkineticscorporation.com/product.php. Accessed February 19, 2016.

[72] Statt N. Intel wearable devices to aid Parkinson's disease research, 2014. http://www.cnet.com/news/intel-wearable-devices-will-help-parkinsons-disease-research/. Accessed January 4, 2016.

[73] Medtronic, Reveal LINQ Insertable Cardiac Monitor, 2015. http://www.medtronic.com/patients/fainting/device/our-insertable-cardiac-monitors/reveal-linq-icm/. Accessed January 4, 2016.

[74] GlySens, Truly Long Term Continuous Glucose Monitoring: The GlySens ICGM® System, 2016. http://glysens.com/products/glysens-icgm/. Accessed January 4, 2016.

[75] Frost M. C., Meyerhoff M. E., Implantable chemical sensors for real-time clinical monitoring: progress and challenges. Curr Opin Chem Biol 6, 633–641 (2002).

[76] Zhang Y., Xiao H., Bluetooth-based sensor networks for remotely monitoring the physiological signals of a patient. IEEE Trans Inf Technol Biomed 13, 1040–1048 (2009).

[77] Park S., Kim J., Ko K., Jang I., Sim K., Real-time heart rate monitoring system based on ring-type pulse oximeter sensor. J Electr Eng Technol 8, 376–384 (2013).

[78] Lee H. J., Lee S. H., Ha K. S., Jang H. C., Chung W. Y., Kim J. Y., Chang Y. S., Yoo D. H., Ubiquitous healthcare service using Zigbee and mobile phone for elderly patients. Int J Med Inform 78, 193–198 (2009).

[79] Ahson S. A., WISP: A Passively Powered UHF RFID Tag with Sensing and Computation. RFID Handbook: Applications, Technology, Security, and Privacy. M. Ilyas, CRC Press, Boca Raton (2008).

[80] Catarinucci L., Colella R., Tarricone L., Sensor Data Transmission Through Passive RFID Tags to Feed Wireless Sensor Networks. 2010 IEEE MTT-S International Microwave Symposium Digest (MTT), Anaheim, CA, May 23–28, (2010), IEEE, Anaheim, CA, pp. 1772–1775.

[81] Mukherjee S., Dolui K., Datta S. K., Patient Health Management System Using e-Health Monitoring Architecture. 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, February 21–22, (2014), IEEE, Gurgaon, pp. 400–405.

[82] Dolui K., Mukherjee S., Datta S. K., Smart Device Sensing Architectures and Applications. 2013 International Computer Science and Engineering Conference (ICSEC), Nakorn Pathom, September 4–6, (2013), IEEE, Nakorn Pathom, pp. 91–96.

[83] Martí R., Delgado J., Security in a Wireless Mobile Health Care System. http://www2.research.att.com/~rjana/Marti-Delgado.pdf. Accessed January 4, 2016.

[84] Bajorek M., Nowak J., The Role of a Mobile Device in a Home Monitoring Healthcare System. 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, September 18–21, (2011), IEEE, Szczecin, pp. 371–374.

[85] Jones V. M., Batista R., Bults R. G. A., op den Akker H., Widya I., Hermens H., Huis in't Veld R., Tonis T., Vollenbroek-Hutten M., Interpreting Streaming Biosignals: in Search of Best Approaches to Augmenting Mobile Health Monitoring with Machine Learning for Adaptive Clinical Decision Support. http://ceur-ws.org/Vol-765/paper3.pdf. Accessed January 4, 2016.

[86] Shahriyar R., Bari F., Kundu G., Ahamed S. I., Akbar M., Intelligent mobile health monitoring system (IMHMS). Int J Control Autom 2, 13–28 (2009).

[87] Abayomi-Alli A., Ikuomola A. J., Aliyu O. A., Abayomi-Alli O., Development of a mobile remote health monitoring system: MRHMS. Afr J Comput Inf Commun Technol 7, 15–22 (2014).

[88] Jones V., Gay V., Leijdekkers P., Body Sensor Networks for Mobile Health Monitoring: Experience in Europe and Australia. ICDS '10. Fourth International Conference on Digital Society, St. Maarten, February 10–16, 2010, IEEE, St. Maarten, pp. 204–209.

[89] Aslantas V., Kurban R., Caglikantar T., A Low-Cost Wireless Healthcare Monitoring System and Communication to a Clinical Alarm Station. http://www.emo.org.tr/ekler/df21078378b5c29_ek.pdf. Accessed January 4, 2016.

[90] Borriello G., Stanford V., Narayanaswami C., Menning W., Pervasive computing in healthcare. IEEE Pervasive Comput 6, 17–19 (2007).

5

COLLABORATIVE OPPORTUNISTIC SENSING OF HUMAN BEHAVIOR WITH MOBILE PHONES

Luis A. Castro[1], Jessica Beltran-Marquez[2], Jesus Favela[3], Edgar Chavez[3], Moises Perez[3], Marcela Rodriguez[4], Rene Navarro[5], and Eduardo Quintana[3]

[1]*Spectrolab, Inc., Sylmar, CA, USA*
[2]*Instituto Tecnológico de Sonora, ITESM, Universidad La Salle, Mexico City, DF, Mexico*
[3]*Department of Computer Sciences, Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico*
[4]*Department of Computer Engineering, Autonomous University of Baja California, Mexicali, BC, Mexico*
[5]*Department of Industrial Engineering, University of Sonora, Hermosillo, SO, Mexico*

5.1 HEALTH AND MOBILE SENSING

Important advances are being made in approaches to infer a variety of contextual information relevant to health care, such as coughing [1], anxiety [2], or socialization [3]. These advances are making possible the use of behavior patterns as the basis for novel pervasive healthcare applications. In this sense, mobile phones are becoming the platform of choice for deploying these applications, given their ubiquity and capabilities. These behavior-aware healthcare applications range from early diagnosis, uncovering lifestyle causes of morbidity, inducing behavior change, disease management, to dealing with problematic behaviors of people suffering from dementia.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

The functional assessment of a patient often relies, at least partially, on self-report, that is, on answers provided by the patient to questions related to his/her habits or symptoms. This approach has validity problems as patients may underreport or exaggerate symptoms. Often, the assessment is based on physical tests performed in a laboratory setting, yet frequent measurements obtained in a naturalistic setting might prove more reliable. For instance, the Timed Up and Go test is frequently used to assess older adults' functional mobility, gait speed, and the risk of falling. Frenken et al. at OFFIS in Germany proposed an approach to perform an equivalent test using ambient sensors in a domestic environment without supervision [4]. The test can be performed continuously, given the physician a more reliable assessment of functional mobility than a test that is performed in a lab every few months, at best. Fatigue is another factor associated with frailty. Unobtrusively monitoring gait speed over a period of time, for instance, in a route frequently followed by an individual, could provide early evidence of fatigue. Similarly, the use of pressure sensors on everyday objects could provide signs of deteriorating grip strength, a predictor of future disability [5].

As an additional example of pervasive healthcare applications based on behavior estimation, consider people suffering from mental disorders who exhibit abnormal behaviors that can put them at risk or burden their relatives and caregivers. It has been estimated, for instance, that up to 90% of patients with Alzheimer's disease exhibit problematic behaviors such as wandering, repetitive questioning, or verbal aggression [6]. Nonpharmacological interventions for dementia aim to ameliorate these behaviors. These, however, are secondary symptoms of the disease; they are often triggered by contextual factors such as changes in the physical environment and social interaction. This is why recommendations for the care of a person with dementia include simplifying the physical environment, avoiding unfamiliar settings, and maintaining a relatively fixed routine. Applications that can take into account behavior could lead to tailored and more opportune interventions by detecting the presence of problematic behaviors and inferring their probable causes, in order to notify the caregiver, offer assurance to the patient, or directly modify the physical environment. For instance, nightfall might trigger confusion and anxiety in persons with dementia, leading them to not recognize their surroundings and a desire to wander—a phenomenon known as sundown syndrome. If this behavior is observed, a common intervention requires increasing the light level, which could eliminate the problematic behavior.

Furthermore, human behavior can often have certain effects on health. Behavioral epidemiology aims at establishing causal relations between behavior and health and proposing interventions to change unhealthy or risky behaviors [7]. Since John Snow inferred that contaminated water was the cause of a cholera outbreak in London by mapping the incidence of cholera and water sources, epidemiology studies have often established unexpected causality between behavior and health. To this aim, proper methods need to be established for measuring behavior. The pervasiveness of mobile phones, coupled with their communication, processing, and sensing capacities, is creating new opportunities for measuring behavior, conducting epidemiological studies and enacting interventions to change behavior. Ubiquitous

sensing technologies can boost epidemiological studies by gathering data from larger populations, increasing the frequency of reporting and providing more reliable data based on the continuous monitoring of actual behavior rather than from sporadic interviews that rely on self-report. Analysis from larger populations, for instance, could result in findings that take into account differences in groups regarding gender, age, or upbringing. Real-time monitoring of behavioral patterns could also lead to early detection of epidemic outbreaks. For instance, an experiment conducted with college students as part of the Reality Mining project at the MIT Media Lab has shown evidence that mobile phone users change their communication habits and have mobility patterns with lower entropy when experiencing symptoms of influenza [8].

The emergence of this fairly new research area known as mobile phone sensing offers unprecedented opportunities for scientists and practitioners to better understand and potentially influence users and contexts. These advances are being helpful in stepping up our understanding of human dynamics and contexts. Even more, a better understanding comes with an opportunity to provide adequate services and potentially influence attitudes and behaviors of people, individually and collectively. However, mobile sensing often faces several challenges that need to be overcome in order to provide meaningful, relevant services. The first challenge is the need for tools to gather fine-grained data of activities and behaviors coming from different sources such as audios, videos, photos, or sensor data. In this regard, smartphones are a ubiquitous, prevalent infrastructure maintained and continually upgraded by the end users themselves. The second issue is understanding the context in which daily life activities are interwoven. Context can be characterized by several variables such as time, identity and roles of people, location, tasks and activities of users, etc. [9]. Estimating some of those variables from sensor data, often sparse and incomplete, is not trivial. There has been a growing demand for more accurate and efficient machine learning and pattern recognition algorithms to infer complex variables such as the current user's activity, mood or attitudes, etc. [3, 10–12]. Last, on top of that, even when there have been advances in mobile hardware, there are still several issues related to the scarcity of resources such as the battery, bandwidth, and storage, which in mobile, battery-driven devices can be central to defining how the system will behave.

In this chapter, we describe some of the challenges associated with the development of pervasive healthcare applications based on the monitoring and inference of behavior among individuals and populations. We present results of two case studies aiming at sensing health-related behavior. The first study centers on the behavior patterns of older adults that might provide evidence of their frailty. The second study aims at detecting problematic or disruptive behaviors of people suffering from dementia, in order to enact proper nonpharmacological interventions to mitigate the adverse effects of these behaviors. Our discussion focuses on the mobile software platforms required to gather the necessary data to make these applications possible, and how collaborative sensing could help address some of the challenges that might prevent mobile phones from continuously sensing behavior, and which are associated with device limitations in terms of battery, storage, and processing capacity. While conducting these studies, we have faced several technical issues related to the limitations of using mobile phones as sensing devices and have proposed approaches to save battery and

data storage, while reducing communication and processing costs. We believe that battery-saving and performance-wise strategies such as the ones we implemented in this work can help other researchers in the area facing similar challenges.

5.2 THE InCense SENSING TOOLKIT

In the past few years, we have been working on designing and implementing a general-purpose mobile phone sensing platform: InCense [13]. Our aim is to offer an open platform that can be used by people with little or no technical background to deploy sensing campaigns. Through a graphical user interface (GUI), InCense facilitates the design and deployment of behavioral data gathering studies from populations of mobile phone users. In this sense, researchers interested in large collections of user context and user input can focus on data analysis stages as well as on the implications of them instead of using their resources in developing platforms to collect the data.

One of the core ideas behind InCense is having loosely coupled components for carrying out particular tasks, which are useful for particular sensing campaigns. For example, knowing if a user is at home, detecting others around, or computing the number of steps a person walks are components that could be independently developed and put together for a particular campaign. These components could be replaced, enhanced, and developed from scratch by other researchers and added to a shared repository. We will describe the high-level details of InCense to explain how the tool was extended into a collaborative application to deal with some of the performance issues faced in mobile sensing, but further details of InCense can be found in Ref. [13].

One of the main advantages of using InCense is that it has low overhead in sensing campaigns that are similar in essence; it runs quietly in the background of a mobile phone; it provides an interface for adding new components; and it can be programmed a priori to react to certain events that occur to the mobile phone (e.g., receiving a phone call) or pertaining to user behavior (e.g., going out). In addition, the InCense platform provides an ontology representing the components as well as relationships among them, which enables us to verify the soundness of a certain campaign. In short, Figure 5.1 shows a high-level description of its architecture and how its components relate to each other.

Among other features, InCense provides a GUI, which facilitates the creation of sensing campaigns. The GUI was designed keeping nontechnical researchers in mind. It provides a series of elements for grouping, processing, and storing unprocessed or preprocessed data. These elements can be dragged, dropped, and connected among them (see Fig. 5.2). InCense provides an independent GUI, as well as an Eclipse plug-in, to design and deploy sensing campaigns. Both GUIs can be separately used to create sensing campaigns, without any difference in the resulting configuration file (i.e., JSON file).

While the features of InCense focus on facilitating the deployment of new sensing campaigns, its performance concerns related to battery, storage, processing, and communication use are shared by similar mobile sensing platforms. The approaches described to address these issues can be potentially applied to these other systems.

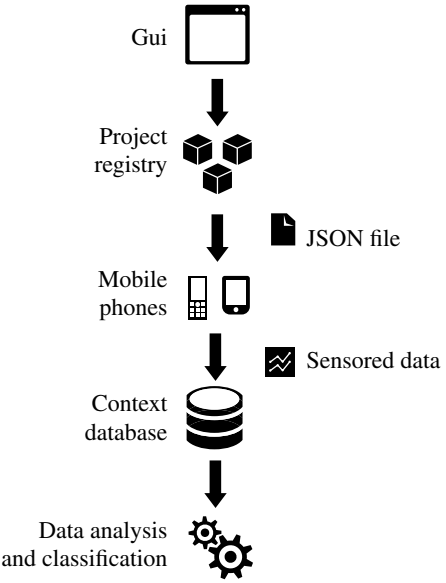


FIGURE 5.1 Architecture of InCense.

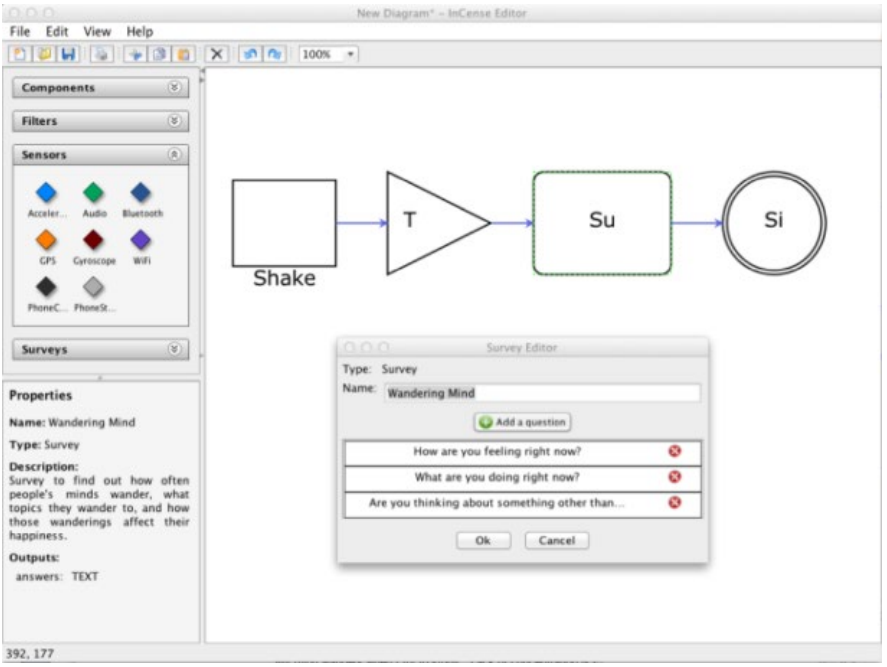
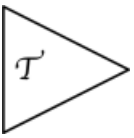
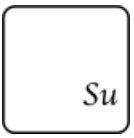
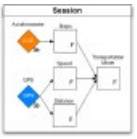


FIGURE 5.2 Snapshot of the independent GUI showing an example where the mobile device collects information through a survey.

5.2.1 InCense Front-End Components

Front-end components are those that are part of the GUI. These components are seen by the researcher or the user and can be combined to create sensing campaigns by dragging them and drawing connections between them in a canvas (see Fig. 5.2). Next, we will present each of these components (Table 5.1).

TABLE 5.1 List of the Front-End Components and Their Corresponding Icon

Icon	Name
	Sensor
	Filter
	Trigger
	Stop
	Survey
	Session
	Sink

5.2.1.1 Sensor Sensors are the source of raw data from the mobile phone built-in sensors. There are different types of sensors such as accelerometer, camera, Bluetooth, Wi-Fi, Global Positioning System (GPS), microphone, and others, and as such each sensor has a particular component implemented.

5.2.1.2 Filter Filters receive input data (e.g., raw sensor data and other's filter output) preprocess, condense, compute, or classify these data, thereby producing higher-level data to facilitate analysis and decision-making. For example, receiving, saving, and transmitting the raw data from accelerometers might be costly. Instead, it may be preferred to use filters to preprocess these data to obtain commonly used signal features such as peaks, standard deviations, and/or means. High-level filters (e.g., human voice detection) can be created through a combination of low-level filters (e.g., fast Fourier transform and Mel-frequency cepstral coefficients or MFCCs). One high-level filter could be used for instance to detect if the user is walking or if it has arrived at a specific location.

5.2.1.3 Trigger Automatic decision-making could be added with triggers. They can receive raw or processed data as input from sensors or filters. When certain conditions are met, they can start collecting data from sensors or launch surveys to elicit user input. For example, if a user has gone out for a walk, a survey is shown to ask how s/he is feeling.

5.2.1.4 Survey They consist of multiple-choice or open-ended questions that the mobile phone users respond to. They can be defined and edited in the GUI like any other component.

5.2.1.5 Stop It halts the execution of a sensing session.

5.2.1.6 Session A session is a group of different components connected to accomplish a particular goal. Sensors, filters, triggers, surveys, sinks, and even other sessions could be added to the workflow of a session. Different sessions could be programmed to be triggered at different times, providing a flexible way of collecting different types of data.

5.2.1.7 Sink They are data pools wherein the information of a sensing session is grouped before being sent to a context database. Sinks can contain raw sensor data (e.g., raw audio files), preprocessed data from filters (e.g., audio fingerprint or AFP), or processed data illustrating human behavior (e.g., the user talked to another person).

5.2.2 InCense Back-End Components

In addition, InCense contains several back-end components that control several aspects that are not necessarily part of the GUI, but still fundamental to sensing, storing, and communicating the sensed or processed data from the mobile device.

The first set of components resides in the mobile nodes. They are in charge of tasks related to communication and data storage as well as keeping InCense up to date.

5.2.2.1 Communication Control It maintains a queue containing the files ready to be submitted to the context database. The designer of the sensing campaign can also decide to transfer the information manually using the USB port of the mobile device. Also, when the network is not available and the device is willing to submit the files, it verifies when the device is back online in order to proceed with the submission.

5.2.2.2 Update Control With some frequency (e.g., once a day), this component can check whether a new version of the mobile application has been released. If that is the case, the new application is downloaded, installed, and run. In addition, this component also checks if a new configuration file for the campaign has been released. If that is the case, the new file is downloaded and loaded so the new settings can be applied as soon as they are received.

5.2.2.3 Storage Control This component checks whether the data to be stored is raw audio, survey data, or other sensor data in tabular form. This helps ensure that the data are saved in the appropriate format.

The following back-end components reside on the InCense server. These services are accessed through a RESTful API.[1] These components are critical to receiving and storing the data coming from the mobile devices.

5.2.2.4 Device Registry This is a database of all devices (mobile and nonmobile) that can potentially participate in a sensing campaign. This registry holds the list of devices and their capabilities (i.e., sensors and processing power). For instance, some devices, such as a local server, might have no sensing capabilities, but can be used for data processing and provide low-latency communication with the mobile nodes.

5.2.2.5 Campaign Registry This registry includes the design of a campaign as well as all the devices that will be participating in it.

5.2.2.6 Context Database This database contains all collected data either in a raw format or processed in some way by any other entity participating (e.g., mobile phone).

5.2.2.7 Context Server This node is responsible for processing the data stored in the context database. Processing might include data analytics to uncover behavioral patterns or correlations of interest to the researcher conducting the sensing campaign, or real-time inferences used by context-aware applications. The server resides in the

[1]REST is a software architectural style, which comprises a series of guidelines and best practices for constructing scalable Web applications. Web applications following the guidelines are said to be RESTful.

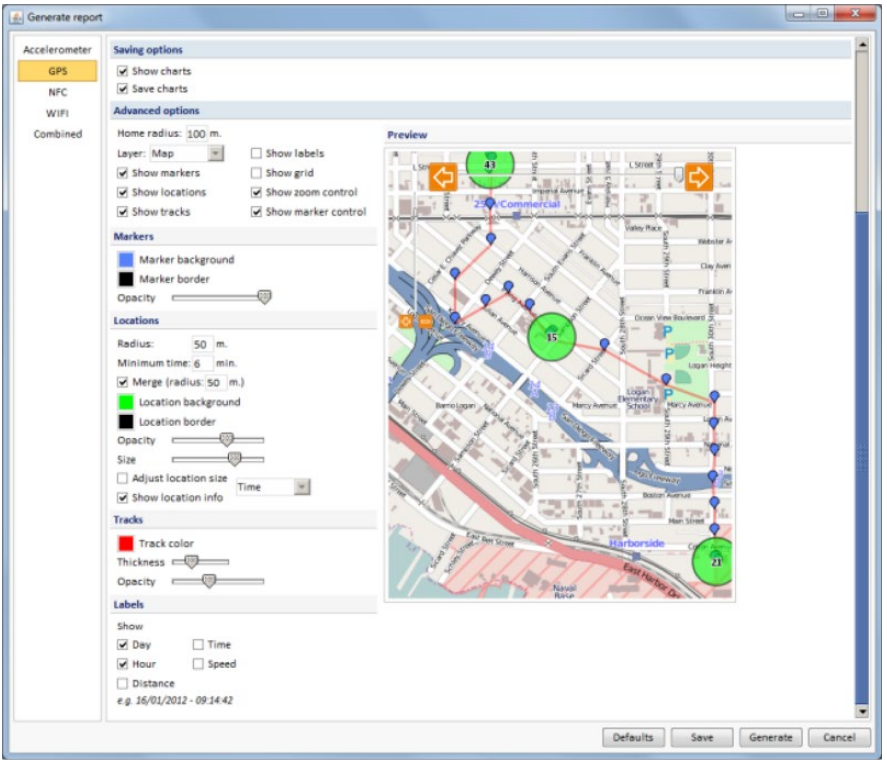


FIGURE 5.3 Visualization and analysis of contextual data from the InCense server.

cloud, providing availability and elasticity to abrupt changes in demand. Figure 5.3 shows a client that accesses the context server to plot behavioral data for further analysis. In this case, GPS data is used to infer the locations frequented by the user.

5.2.3 Collaborative Sensing with InCense

Due to the nature of the sensing campaigns, some of the components are fundamentally mobile and reside in the mobile phones, whereas others reside in nonmobile nodes such as a local server or in the cloud. This sole feature adds significant complexity to the design of the campaigns as the researcher has to somehow make several considerations a priori, if resources are to be properly utilized, such as the processing and data transfer workload on mobile phones as well as servers. This configuration is not trivial as it involves considering multiple variables related to the diversity of the participating devices such as the processing power, battery consumption, number of devices involved, types of data to be sensed, server location, and data transfer rates and costs. Moreover, several of those variables can vary from one moment to another (e.g., number of available memory and battery left) and have to be monitored in real time. In order to intrinsically support and consider some of these aspects since the

inception of a sensing campaign, several characteristics have been added to a collaborative version of InCense, which are next discussed.

Mobile and nonmobile devices participating in an InCense sensing campaign can perform different tasks that can be classified into four types: (i) sensing, (ii) processing, (iii) communication, and (iv) storage. In this regard, based on different conditions such as battery remaining, processing power of each participating device, communication costs, etc., InCense could illustrate trade-offs among these aspects to account for them at sensing campaign design stages. For instance, mobile nodes usually have sensing capabilities and limited processing power and storage; on the other hand, servers in the cloud provide advanced processing power, massive storage, but high latency. Choosing one or another depends very much on the performance of each of them. To illustrate this, consider an application that needs to process high-quality images to look for certain features such as points of interest, and then search among similar images in a database. This task demands considerable processing power (and battery). In this case, a better strategy could be having a local server performing an image analysis, as it may be computationally cheaper to send it over the network.

5.2.3.1 Collaborative Modalities with InCense We extended InCense to support collaborative modalities that can help increase the performance of the system by using idle and/or better-suited devices for particular tasks as well as increase the battery life of the mobile devices involved. We next describe each of these modalities based on the sample InCense app in Figure 5.4.

Collaborative Processing When some of the tasks are too demanding for a particular device, one or more tasks could be delegated to other better-suited devices, such as those that might be idle. For instance, when two or more mobile devices are nearby and sensing, some of the processing can be performed in only one of them. Also, if available, some heavy processing can be performed on a local server (Fig. 5.5).

Collaborative Sensing When two or more devices are nearby, there is no need to sense and save this information in all the devices . Instead, one device can sense one particular aspect while others can be in charge of sensing other aspects, meaning

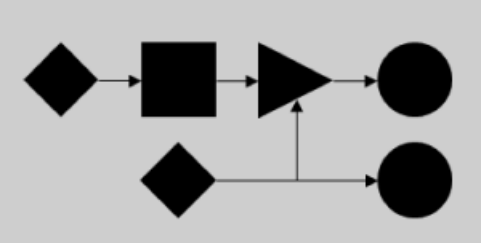


FIGURE 5.4 Original InCense app. Two sensors (diamond = sensor), one directly saves the info (circle = sink) and the second one has a filter (rectangle) and based on the value of the other sensor decides whether to record or not.

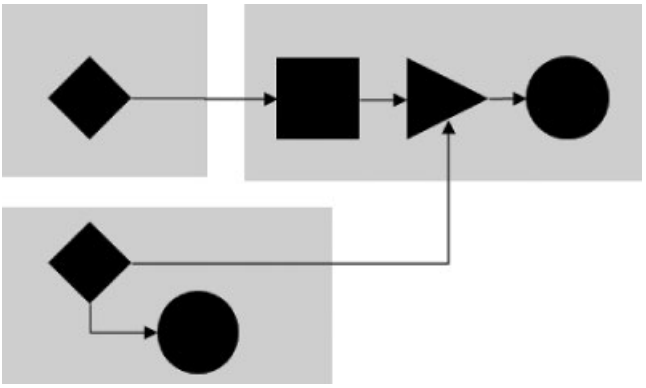


FIGURE 5.5 One of the devices only senses information; an additional device (below) not only senses and saves the data in a Sink but also sends the data to another device (a local server), which receives data, performs some computations, and then it decides whether to save it or not.

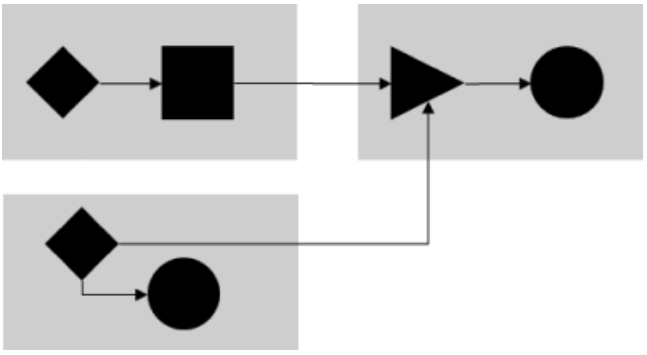


FIGURE 5.6 Two devices (left) are sensing different data and submit the data (preprocessed and raw data) to a third device, which will decide whether to save it or not.

that all devices can split the sensing tasks, saving resources such as battery and storage (Fig. 5.6).

Collaborative Storage Depending on the characteristics of the devices, large chunks of collected data such as audio or video can be stored on a device with larger storage capacity such as a local server (Fig. 5.7).

Collaborative Transmission Some devices are better suited for particular tasks since they have different characteristics. If available, mobile nodes can locally transfer data to a local server for its transmission using a broadband. Also, mobile nodes can split the submission of data to a server so they can save battery and on data services costs.

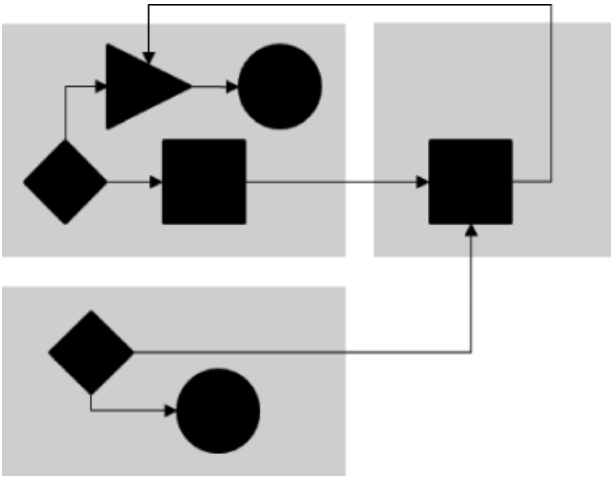


FIGURE 5.7 Two devices are sensing data, and a third device only processes data (a local server or in the cloud); the results are used by one of the devices to decide whether to save it or discard it.

Task Allocation Based on Device's Capabilities, Resources, and Cost All these collaborative strategies have to be accompanied by algorithms that enable optimal or near-optimal task allocation. We have been implementing some basic algorithms based on heuristics, but advanced, automated algorithms for allocation must account for the following aspects.

First, the Device Registry and the Campaign Registry have data about all the devices participating in a particular sensing campaign. These data can be used to allocate tasks that suit well a particular device in terms of its capabilities. Participating devices can be very diverse in terms of processing power, memory, battery, storage, data transfer rates, latency, and the like. Therefore, taking into account these aspects not only at the design stage but also at runtime can enhance the performance of the overall system.

Second, besides the list of devices and their capabilities (i.e., resources), these algorithms have to consider current and future availability of those resources. That is, for instance, one of the best mobile devices is not particularly useful if it is busy with other processing tasks (i.e., watching a live streaming video). Therefore, the devices must continually report or be monitored regarding aspects such as credit (i.e., prepaid phones), memory available for carrying out tasks, number of tasks lined up, remaining battery, signal strength, location of the user, etc. Moreover, algorithms could consider not only current availability but also future availability based on the previous history of behavior (e.g., every night the phone is idle, so it can perform costly processing tasks). Advanced task allocation algorithms could consider these aspects and be somewhat successful, in so far as these aspects can be monitored or supervised in real time.

A third aspect to consider is computational cost. Some processing, sensing, storage, or transmission can be more costly than others. Therefore, the cost is a factor that has to be considered as well. In this regard, considerations such as having on-device versus server processing, local versus nonlocal processing, and sending data now or at the end of the day are fundamental and must be accounted for. In the example presented in Section 5.2.3, the image feature extraction could be carried on the mobile phone, but it is often too costly in terms of battery as well as processing power. Sending images to the local network and having a local processor carry out the heavy processing tasks can be a good compromise between latency and overall performance. Still, as mentioned already, some of the algorithms we used were implemented a posteriori, but more advanced algorithms can take advantage of optimization strategies as well as task parallelism.

The previous sections illustrated how collaborative modalities can account for the multiplicity of data that can be generated and sensed at particular locations. In the following sections, we describe two sensing campaigns aimed at gathering behavioral data from older adults. We describe the design of the InCense sensing script for both campaigns, the performance issues raised in each of them, and how they were addressed.

5.3 SENSING CAMPAIGN 1: DETECTING BEHAVIORS ASSOCIATED WITH THE FRAILTY SYNDROME AMONG OLDER ADULTS

We designed a sensing campaign aimed at detecting behaviors commonly associated with frailty in older adults. Frailty is a state of increased vulnerability to adverse health outcomes for people of the same age. The occurrence of frailty increases with advancing age. As such, frailty is a major challenge for the healthcare systems as frail individuals have higher mortality rates and require more health care than people who are fit [14]. Frail people are at high risk for major adverse health outcomes, including disability, falls, institutionalization, hospitalization, and mortality. The clinical assessment of older adults is to a large extent based on retrospective accounts of incidents. This can be unreliable as patients often do not remember or try to hide or minimize negative incidents. For example, widely accepted surveys to estimate frailty in older patients include questions such as “In the last week, how many days did you walk at least 10 minutes?” and “How frequently do you speak with your friends/spouse?” Often, responses to these questions are hardly precise, having older adults providing rather vague answers to questions of this nature. Therefore, we aimed at sensing and estimating some of those variables (e.g., physical activity and socializing with others) that could correlate to surveyed data of older adults pertaining to frailty.

The sensing campaign was aimed at collecting data to assess frailty in older adults and compare it with the results of the clinical assessment. The functional assessment of frailty in elders for the campaign included a clinical assessment by a qualified physician who applied a baseline and exit questionnaire that included standard inventories such as the ADL-KATZ [15], the SF-36 health survey [16], the Mini Nutritional

Assessment (MNA) [17], and the International Physical Activity Questionnaire (IPAQ) [18, 19]. The devices used were Samsung Nexus S smartphones running Android Gingerbread (2.3). However, InCense can run in lower-end devices as well as in newer versions of Android.[2]

We recruited 15 community-dwelling older adults to carry the mobile phones in order to gather data from their activities and behaviors. The participants' average age was 75.3 (SD = 1.8), ranging from 73 to 79 years old. In addition to the mobile devices with InCense, older adults were provided with a deck of 20 cards depicting some of their most common activities such as cooking, watching TV, going to the doctor, and the like. These cards were enabled with near-field communication (NFC) tags, which we used to obtain the ground truth. After each activity registered through the NFC tags, 2 min of audio were recorded. Even when InCense allowed the mobile device to run for a couple of days without the battery running out, they were instructed to charge the phone every night. In case of a device reboot, InCense would automatically restart by itself.

We collected data from several sources in the mobile phones including accelerometer, GPS, Wi-Fi, NFC tags, as well as the microphone. The data collected from all participants sum up to around 47 GB (compressed ~17 GB), having around 3.1 GB per participant. Every week, the participants received a visit from two members of the research team to check on them, receive feedback, and obtain the data from the mobile devices using the USB port. Given that older adults did not have access to the network in their homes, and due to the costs associated with sending data through the global system for mobile communication (GSM) data services, data were collected in situ. All collected data were analyzed a posteriori in the lab.

5.3.1 InCense Sensing Script for Detecting Behaviors Associated with the Frailty Syndrome

The InCense configuration for this sensing campaign is presented in Figure 5.8. This InCense app is in charge of collecting raw behavioral data from older adults. In the upper section of Figure 5.8, a trigger waits for an NFC tag to be detected; when this happens, it saves the tag label (i.e., brushing teeth) in a data Sink, along with 2 min of audio in an audio Sink. All recorded data are time stamped. Simultaneously, the mobile phone is continuously capturing accelerometer data and saving these in a data Sink. In the lower section of Figure 5.8, a connection filter is continuously checking whether a home wireless access point is available. When the access point is detected, meaning that the user is at home, a trigger stops the GPS sampling to conserve the battery. When the home network is not detected, the trigger reinitiates the GPS sampling, meaning that the user has gone out.

5.3.2 Results

The frailty syndrome includes assessing the physical conditions of the individual in terms of sarcopenia, osteoporosis, and muscle weakness. Some of them are measured directly, whereas others are measured indirectly (e.g., walking). Based on the frailty

[2]InCense has been used in devices running Android KitKat (4.4.2).

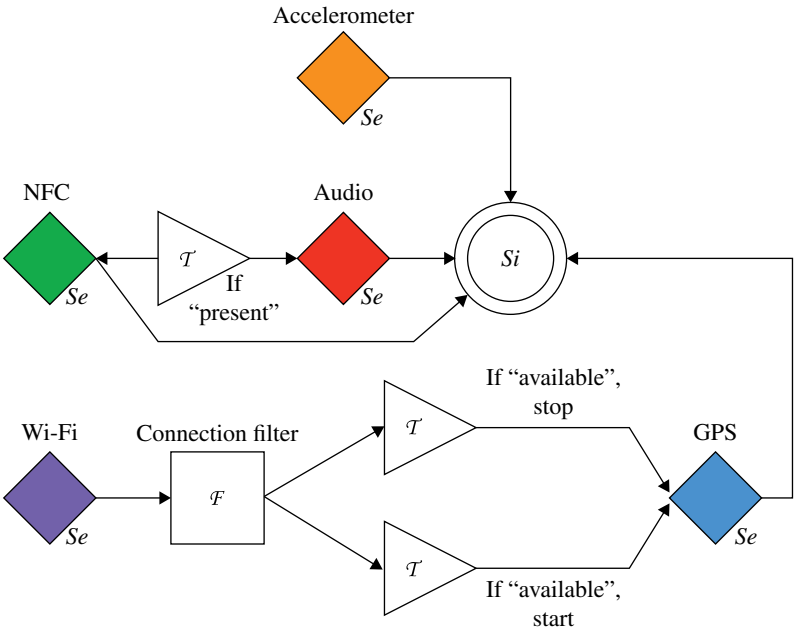


FIGURE 5.8 InCense app for detecting behaviors associated with the frailty syndrome in older adults.

index developed by Fried [20], four participants were classified as frail during the baseline interviews.

Taking this classification as the departure point, we made some links between observed behaviors and frail/fit participants. From our data, frailty seems to have some adverse effects on participants. Some bivariate correlation coefficients in frail adults that were of interest are depression and sleeping time, as well as depression and going out of the household. Although not very surprising, our data shows that the sleeping time in frail participants was larger than for fit participants. Also, frail participants went out of their households less often than fit participants. One benefit of using sensor data is that, even when they are error prone, they tend to be more reliable than self-report. The complete results of this case study can be found in Ref. [13].

In terms of physical activity, it is important as it helps adult muscles grow stronger, thus monitoring the physical activity can be important to encourage exercise or raise the alarm when physical activity decreases. In adults aged 65 years and above, physical activity includes walking, dancing, gardening, household chores, play, games, sports, or planned exercise. From our data, we found that participants who were fit performed significantly more intense activity bouts than those who were frail.

Mobility in older adults is also important as it can be used to determine an individual's life space, which can be used to identify individuals in a community that may be at risk of not living a prosperous life [21]. The distance from the home

location can be used to estimate the individual's geographic life space [22]. This is important since a decline in the life space may indicate early functional decline in older adults [23]. In this case, therefore, having the mobile phone collect this sort of information can be essential for better diagnoses as most patient assessments are performed on the basis of self-reports.

5.3.3 Performance Improvement During the Sensing Campaign

This section presents some of the performance issues we encountered during this campaign and how they were addressed. In particular, we will discuss issues related to battery draining processes, performance enhancements, and other battery-saving strategies.

5.3.3.1 Issues with the Battery In the beginning, we experienced several issues with the battery. Within 4 h or less, the battery was completely depleted. For obvious reasons, this was not practical at all as older adults were under no supervision and they would have to be continually charging their mobile phones.

After carefully reviewing the code, and conducting some experiments, we decided to reduce the sensitivity (from maximum to medium) and the intervals at which the accelerometer was sampling. Less accelerometer sensitivity meant that we lost some details of the signal, but it was still effective for computing the amount of physical activity of the phone bearer.

5.3.3.2 Performance Improvement We conducted an experiment to test the battery consumption while sensing accelerometer data. We programmed two Android apps that monitor accelerometer data and compute activity counts from 1 min epochs. We requested and stored the battery level every 10 min. In order to save the battery, one app enters in stand-by mode (30-s window) when the activity counts are below a given threshold. Then, checks for movements (3 s), if no movement is detected, the app returns to stand-by mode. These apps were installed onto two Samsung Nexus S smartphones running Android Jelly Bean (4.1.2). One participant carried these phones (in the same pocket) for 20 consecutive hours. Figure 5.9 shows the changes in battery level using both approaches. We observed a 20% battery life saving when finding sleep opportunities. The 30-s stand-by window provides a good compromise between data loss and battery life.

5.3.3.3 Multisource Strategies As commented throughout this chapter, one of the main issues in mobile computing is the battery life. Thus, implementing strategies that can help reduce battery consumption can reduce load on the user by not having to charge the mobile phone every few hours and improve our chances of capturing meaningful data.

An additional strategy that we implemented in this campaign was reducing the GPS sampling since this is a battery-depleting sensor. Therefore, as a strategy to save the battery, we also provided a Wi-Fi access point with a known service set identifier (SSID) (i.e., network name). Detecting this SSID meant that the user was at home.

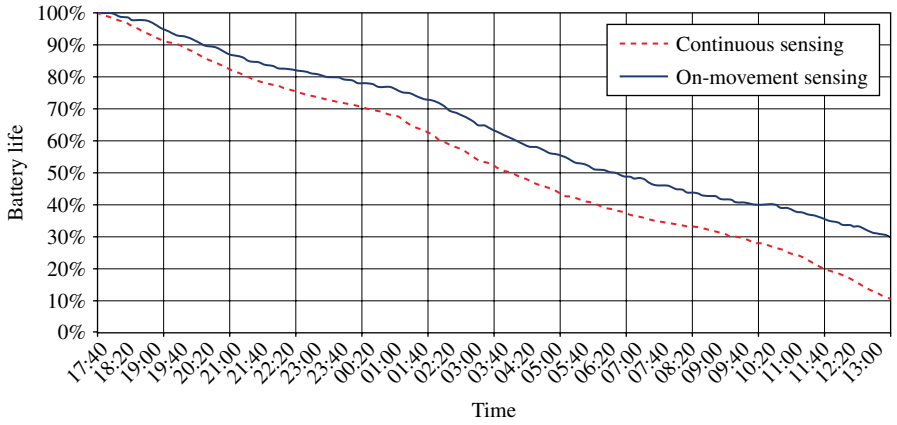


FIGURE 5.9 Savings in battery life when implementing a "stand-by" approach to accelerometer data sensing.

After some consecutive readings without detecting the home network (our Wi-Fi access point), we assumed that the older adult had gone out and turned on the GPS sampling. On the contrary, when detecting the home network, the GPS sampling was turned off. Occasionally, the mobile phones got intermittently disconnected from the wireless network, resulting in having several GPS readings at the home location, which were easily identified.

This proved effective since older adults normally spend a considerable amount of time at home. In our study, participants remained at home for more than 22 h a day on average, with a couple of them seldom leaving home during the whole 3 weeks of the study. This is consistent with previous studies on the mobility of older adults.

Collecting GPS samples all day (once every minute) requires 320 kB of space, which do not represent a real concern in the short term, either for storing or sending data over the network. However, when a sensing campaign has many participants, every byte is important. Turning on the GPS only when the user is outdoors can save around 320 kB a day. During the sensing campaign with older adults, we collected 10.8 MB of GPS data, almost a tenth of what would be required if the GPS records data all day.

These changes for sensing GPS and accelerometer data increased the battery life from 4 h to more than 16 h. This was indeed effective since older adults could carry the mobile phones during the day, and recharge their phones at night, thus making our sensing campaign feasible.

5.4 SENSING CAMPAIGN 2: DETECTING PROBLEMATIC BEHAVIORS AMONG ELDERS WITH DEMENTIA

We designed an observational study of problematic behaviors in institutionalized persons with dementia (PwD). The aim of the study was to identify behavioral and psychological symptoms of dementia (BPSD) such as agitation, anxiety, and

depression in order to implement a suitable intervention to support the PwD and their caregiver. These behaviors are characterized by different physiological signs such as repetitive body movements, variations in walking gait, shouting, and/or crying. Thus, detecting these movements through the built-in sensors on mobile phones seems feasible since those bodily expressions are usually very pronounced. In collaboration with the staff of a geriatric home, 7 residents (Females = 5, Males = 2, ranging from 81 to 94 years old) were selected to participate in the study. During 2 weeks for 4 h a day, participants were closely observed to document BPSD incidents using BPSDiary, an Android application based in the Neuropsychiatric Inventory Questionnaire (NPI-Q), a widely accepted clinical instrument for evaluating psychopathology in dementia. The distress in the PwD caused by the BPSD is usually rated as mild, moderate, or severe. Precise signs exhibited by the PwD, along with the duration of the episode, were recorded for each incident. Additionally, we kept track of social interaction with the residence staff and other residents. Throughout the observation, subjects were carrying a mobile phone running InCense to collect accelerometer data (e.g., to identify physiological manifestations such as repetitive motion and wandering) and ambient sound (e.g., to detect shouting and/or repetitive verbalizations). Once an incident was recorded, the BPSDiary application recommended to the caregiver a suitable intervention based on the actual context and the profile of the PwD.

Additionally, a collaborative sensing module of auditory information was used with the purpose of reducing the audio processing by detecting proximity of two or more subjects. When the devices carried by the subjects are close enough, the audio in both devices is very similar, and the module can decide which device is more appropriate to record and process the audio based on its current status of battery, memory, and CPU usage. Proximity between subjects can also be a valuable data in itself because it provides information about context and socialization.

5.4.1 Detecting Problematic Behaviors in Older Adults

As suggested, we need to be able to automatically detect specific types of disruptive audible manifestations (DAM) [4]. After an analysis of the recordings from an inventory of disruptive sounds from all residents, we proposed the system depicted in Figure 5.10. We generated models for three main types of sounds. The first type are keywords (e.g., *shut up, silencio, and nurse*) frequently uttered by some of the participants, some of them in English and some in Spanish. The second type is disruptive vocalization, like mumbling and other utterances frequently exhibited by residents. The last type corresponds to specific environmental sounds like the tapping on a wheelchair produced by a participant when he/she is agitated and wants to leave the house, or the sound of flushing the toilet. The input audio is available to all the models in parallel.

We made use of techniques developed for speech [24] and environmental sound recognition [25] to detect DAMs. The typical procedure for sound recognition consists of representing the audio as a vector of features obtained through feature extraction. This is followed by a classification stage, which indicates if the audio corresponds to a sound type in a previously trained model.

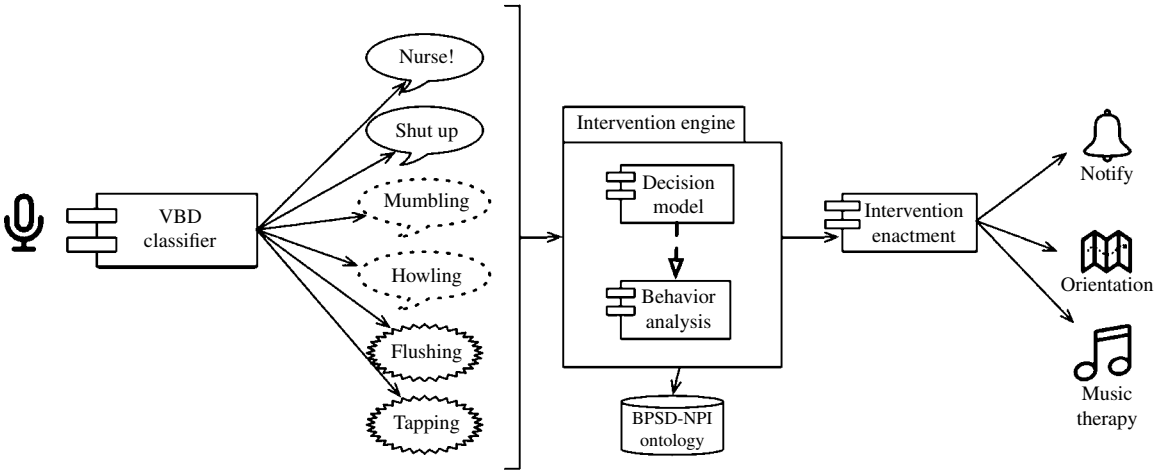


FIGURE 5.10 Ambient-assisted intervention system based on the detection of disruptive auditory manifestations and the enactment of proper nonpharmacological interventions.

Each keyword model is constructed using examples of the word to be identified. Those examples will come from a stream of sounds recorded from the residents suffering from dementia. We have decided to use MFCCs[3] as audio features, since it is the state-of-the-art feature set for speech processing.

In addition, hidden Markov models (HMM)s have been found to be appropriate to handle the regularity in patterns proper of human speech. From the data collected from one of the participants (P3) in the field study, we collected 19 examples of the keyword *shut up*, 12 examples of *silencio*, 9 examples of *chicken shit*, and 3 examples of *stop it*. From the total 60 h recorded from the P3, we found that, in average, he/she says 0.7 of these keywords per hour. In particular, we generated models for the keywords *shut up* and *silencio*. Then, we evaluated them using threefold cross validation obtaining an F1 score of 91%.

A second type of DAM is mumbling. In the literature, there are works looking at detecting baby babble [26], but none were found looking particularly at DAMs in elders. From our data, we analyzed 113 h of audio from all participants, and we found 98 instances of DAMs, mainly from P1, having a rate of 1.4 DAMs/h.

We obtained results from an experiment for classifying mumbling. The database is formed by manually segmenting 43 instances of mumbling recorded from a participant in the field study. We used additionally 43 manually cut examples of random environmental data also captured in the field study. On average, the length of each instance is 2.25 s. The feature extraction techniques used to represent each sound are MFCC, spectral roll off, spectral flux, spectral flatness, and spectral centroid. The classifier used for this experiment is a continuous ergodic HMM with one Gaussian and three states. We evaluated the classification using 10-fold cross validation, obtaining the confusion matrix depicted in Table 5.2. Confusion matrices or contingency tables are widely used in machine learning and usually display the performance of a machine-learning algorithm. Following Table 5.2, rows represent the actual classes whereas columns represent estimations. To illustrate this, from Table 5.2, *mumbling* was, on average, correctly estimated 81.24% across all examinations; whereas in 18.75% of the examinations, they were inaccurately classified as *other* when in fact they should have been classified as mumbling (i.e., false negatives). As seen, false positives for mumbling were more frequent, meaning that the algorithm incorrectly identified *other* sounds as *mumbling*.

We also experimented with 6 min of continuous audio streams recorded in the residence containing 33 instances of mumbling in which we obtained an F1 score of 87%.

TABLE 5.2 Confusion Matrix for Mumbling and Other Sounds

	Mumbling (%)	Other (%)
Mumbling	81.24	18.75
Other	38.28	61.17

[3]The MFCCs are coefficients based on the Mel scale, which is related to human hearing [1]. To calculate the MFCC, it is necessary to analyze the signal in short time, overlapped windows to obtain the cosine transform of a log power spectrum from the Mel-scale frequency bands [2].

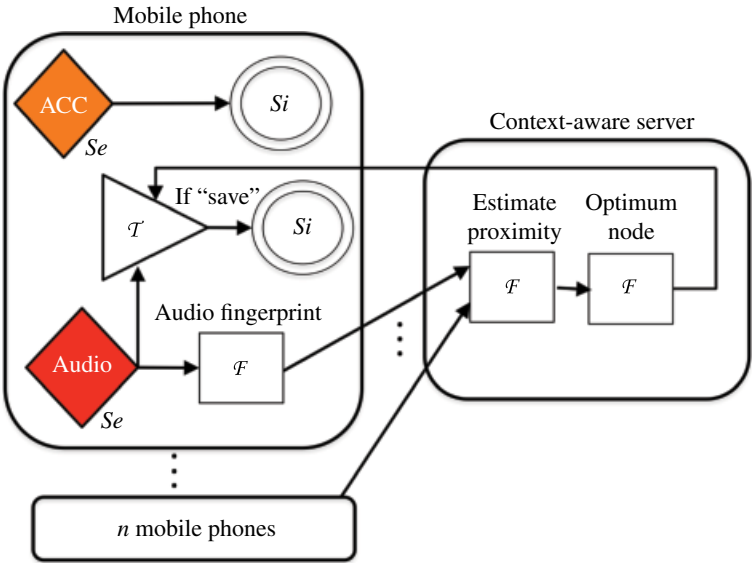


FIGURE 5.11 InCense app for collaborative auditory sensing.

For the type of DAM related to environmental sounds, we proposed an approach that detects environmental sounds such as washing hands or brushing teeth, which can be good enough for certain applications for an assisted living [27]. Some adjustments in the feature selection will be needed to detect tapping on the wheelchair, or the sound of flushing a toilet.

5.4.2 Performance Improvement During the Sensing Campaign

This section presents some of the performance issues we faced when preparing the sensing campaign and how they were addressed. In particular, we present the strategies that were used to reduce the processing and storage when two or more devices are nearby.

5.4.2.1 InCense Sensing Script for Detecting Physical Proximity of Two Subjects Figure 5.11 shows the InCense configuration for this sensing campaign. This InCense app is divided into two sections: one runs on each of the participant's mobile phones and the other on a server located in the nursing home where behavioral data gathering was conducted. In the mobile app, an InCense sensor captures raw data from the accelerometer to be stored in a data Sink for posterior analysis. Additionally, a collaborative sensing module is used to decide which mobile phone is more appropriate to capture audio when two or more subjects are close enough to record a very similar audio signal. In addition, each mobile phone captures audio raw data that is fed into a filter to extract an AFP, which is a content-based audio representation that can be used to measure the similarity between audio signals [28].

To this end, the "estimate proximity" filter, residing on the server, receives AFPs from all mobile phones to identify in those recordings fairly similar audio information. Finally, the "optimum node" is a filter that decides which mobile phone will carry out the audio recording and then triggers a condition to deactivate recording in those mobile phones not selected, thus implementing a collaborative sensing strategy. As commented in Section 5.2, component composition in InCense is loosely coupled, thus new algorithms can be added by just replacing the filter for the "optimum node," which in fact should consider multiple devices. As it is, the node that will continue recording is selected at random, but more sophisticated strategies that take into account the battery life, storage capacity, and the bandwidth at which the node is currently transmitting could be taken into consideration to make the decision.

5.4.2.2 Reducing Audio Processing and Storage The goal of this experiment is to use auditory information to detect devices that are recording the same audio information. The setup is twofold, on the one hand, we can optimize computing and resource usage in the devices. On the other hand, shared audio cues hint proximity between subjects, and this information can be valuable sensing data in itself.

Raw audio data is both large and sensitive to noise; hence, we use a compact representation of the audio called AFP. This serves two purposes, on the one hand, it reduces the transmission overhead to the context server, and, on the other hand, being a stable representation of the audio it can also be used to discover if two devices are recording the same scene. It has been shown in Ref. [28] that computing just the amount of information carried by a signal is robust under changes of the amount of energy in the signal, low-pass filtering, compression, and small amounts of white noise. To compute this AFP, we first obtain a vector with the entropies from the incoming audio signal. For this, we calculate the entropy every frame of size approximately 370 ms with an overlap of 87.5%. Then, the entropy vector is binary encoded to reflect only the modulation of the entropy over time—for example, if the entropy increases from one frame to the next, it is encoded as 1, if it stays the same or decreases, it is encoded as a 0. This AFP is very compact using only 1 bit per frame. Two AFP can be compared using the Hamming distance.[4] For a WAV audio file of 6 min, the size of the AFP is 0.012% of the original size, and it takes approximately 540 ms to process 6 min of an audio. An advantage of this AFP is that we do not send the original audio to the server, and since we cannot inverse transform the data, we maintain the privacy of the users.

To test our method, we collected two sets of data—the test data 1 was recorded in the lab, while the test data 2 was recorded in the older adults' residence. Both sets were captured with a 24,000 Hz sampling frequency with 8 bits of depth. The test data 1 consists of 18 min of audio collected with three recording devices (i.e., A, B, and C). In this scenario, we can observe the following combinations of data recordings: A, B, and C are together, A and B are together but not C, A and C are together but not B, B and C are together but not A, and finally, all the devices are in different

[4]The Hamming distance between two vectors of equal length is the number of coefficients in which they differ.

locations and therefore recording different data. The test data 2 consists in 6 min 30 s of audio collected with two recording devices (i.e., A and B). In this scenario, we can observe the following combinations of data recordings: A and B are together, and A and B are at different locations.

The support for comparing two audio streams is 15 s. We compare every 15-s window of one stream with the corresponding neighboring windows of the second stream up to a shift threshold of 300 ms. This shift accounts for a time lapse or phase shift caused by the audio arriving to the recording devices from different relative positions. For each possible alignment we compute the corresponding Hamming distance and locate the minimum. Since we are overlapping the measures, every distance computation represents 1.7 s of an audio.

After this preprocessing, we classify the stream just as joint/disjoint. We set by hand two thresholds (upper and lower) for the distance; if something is in between, it remains unclassified in a first pass. On a second pass, we get rid of small flips in the class value if the time support is smaller than a time threshold.

We tested two pairs of thresholds: TH1 and TH2. Figure 5.12a corresponds to the comparison of A and B of the test data 1, and Figure 5.12b corresponds to the comparison of A and B of the test data 2. The D row shows the distances, where a darker color represents similitude (i.e., devices are close), and a lighter color represents the difference (i.e., devices are at different locations). The GT row corresponds

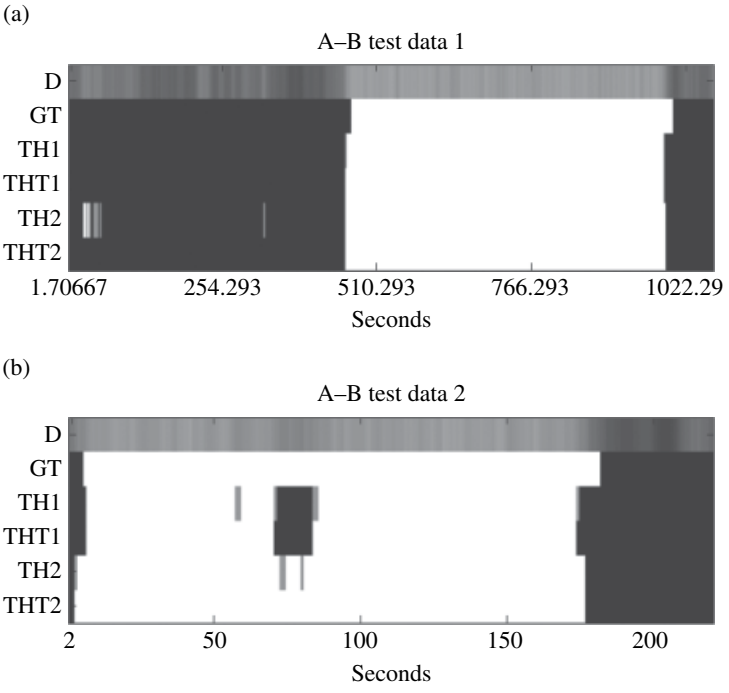


FIGURE 5.12 Results comparing devices A–B: (a) test data 1 and (b) test data 2.

TABLE 5.3 Accuracy and False Positives in Test Data 1 and Test Data 2

		Accuracy (%)		False Negatives (%)	
	Recording Devices	TH1	TH2	TH1	TH2
Test data 1	A–B	97.92	98.08	2.57	1.92
	A–C	94.37	84.20	3.11	1.91
	B–C	98.08	97.35	1.43	0.95
Test data 2	A–B	93.75	95.17	12.5	2.84

to the ground truth; here the black color means that both devices are together and white means the opposite. The TH1 and TH2 rows correspond to the classification using the two sets of thresholds, here black color refers to a prediction that both devices are together, white is the opposite, and the gray color represents the undetermined classifications. The THT1 and THT2 rows correspond to using a time threshold on the results of TH1 and TH2 to eliminate the flips.

In Table 5.3, we show the accuracy and false negatives of the results using the time thresholds on both TH1 and TH2. The approach achieves an accuracy of over 90% due mostly to false negatives in the boundaries. False positives are preferred over false negatives since we value a complete recording over power savings, hence they should be kept to a minimum. False negatives mainly coincide at the frontiers of devices approaching or separating. The source of error is the time discretization (we used a sliding window of 15 s, which can be reduced to obtain a more accurate prediction at the frontier). For this, we suggest not to ask the device to stop recording until we have detected a given amount of time in which both devices are nearby.

One of the main challenges encountered in this experiment is the delay of the incoming audio in the devices due to the positions of the users. Also the reverberation produced in enclosed spaces. An additional challenge was encountered in the older adults' residence, where rooms are connected by a hallway and some audio information could be listened in more than one location. In other words, a shared source of audio (e.g., one person standing in the doorway) may induce a false link between two devices recording in separate rooms. Moreover, two residents may be watching the same TV show in separate rooms and a false audio link may be deducted. If they are sending information to the sensing server from different access points, the ambiguity may be resolved; on the contrary, this issue can be more complex if they share the access point. Local information about the audio topology of the recording environment may help in the design of a disambiguation method, fingerprinting the audio of popular TV shows may help as well. Similar work has been done with an AFP based on the MFCCs [29] or the recently used multiband spectral entropy [30]. Those AFPs require more computing power; hence, there is a need to avoid duplicate recording to save both bandwidth and computing power, which translate to battery savings. To give an example, if we were using the MFCC AFP, every hour of recording produces about 105 MB of data; hence, it would be a significant saving in bandwidth.

The multiband spectral entropy AFP has a smaller footprint (6.17 MB), nevertheless the computing power used could also be significant.

The data gathered can be useful to the researchers because along with labels of the activities or behaviors it can be used to encounter patterns that could be used to design filters that automatically detect these activities or high-level behaviors, thus enabling the selection of the more suitable intervention.

5.5 DISCUSSION

When resources are limited, they need to be regulated in order to help optimize their utilization. This is paramount in environments where resources are particularly scarce. In mobile computing, one of the most valued resources is battery life. Therefore, some strategies or alternative algorithms need to be employed should we need to save battery (see Ref. [31]) and increase the performance of the overall system. Throughout this chapter, we have been discussing how, based on actual sensing campaigns, we look to optimize the use of the processing power and storage in order to minimize battery depletion.

In order to step up most of these strategies, more complex algorithms are needed to make automated, near-optimal decisions during runtime. For instance, the decision whether to process a particular task on a mobile device versus a server depends on a composite interplay of factors that contributes to the complexity of taking these sorts of decisions a priori such as network traffic, current processing load, and the like. These types of decisions are not trivial and can have a great impact on the overall performance of the system. Therefore, they merit careful consideration—and several tests—before implementation since in many cases these decisions can be expensive and detrimental to the functioning of the system. On the other hand, having a way to sense not only the external variables surrounding the devices but also its internal status, such as the memory available, allocated tasks, battery life, signal strength, and others, can open the door for far more complex algorithms. One such algorithm is selective task allocation, in which device capabilities, current and future internal states, and cost-oriented factors could help decide which device to utilize in real time. For instance, having three mobile devices nearby sensing more or less the same information could result in the following configuration: one mobile device senses all the information, a second with a better processor and plenty of battery is in charge of preprocessing data, and a third will send the data over the network. However, different conditions could have resulted in a different configuration.

Moreover, having two or more devices locally sharing the data that have been generated elsewhere can be particularly useful when the generated data involve millions of computations derived from several multidimensional matrixes (for instance). Much simpler strategies can be applied such as sharing location with other devices in the vicinity either to save battery or because some devices do not have a GPS, but still can obtain their location through Bluetooth. Similar strategies can be employed also to reduce uncertainty, that is, when a device is not sure it “heard” (i.e., sensed) something, it could check with a nearby device. Our approach to avoid duplicate

recording of audio could be used either to stop further processing or to deter a device from sensing the audio over the network, that is, more than one device is listening, but only one classifies and sends the results to a server.

Without any doubts, there is plenty of work to do in this regard as distributed, simultaneous mobile sensing is increasingly becoming important due to the pervasiveness of mobile phones and their increasing sensing and processing capabilities. Much more efficient classification algorithms are needed or a combination of strategies that can help us increase their accuracy (e.g., see Ref. [11]).

Furthermore, more development is needed if researchers and practitioners are to use these data. As is, the InCense platform is still limited in readily providing a friendly environment through which data can be collected, analyzed, and interpreted by practitioners. There is plenty of work ahead in developing tools that can work on top of the technical parts that we have described in this chapter, which hopefully are not any harder than, say, driving a car (without having to manually connect wires, plugs, and sensors, and not having to worry about how the engine works internally).

As of now, the InCense platform can be used by the research community to collect data, analyze them, draw some interesting results, and pose new questions. However, there is an implicit serious overhead involving several hours for analyzing the data in the lab. Once this work is carried out, and perhaps some interesting behavior–health connections are drawn, the InCense platform (or any other similar platform) can be used for supporting the work of physicians at their offices by incorporating data gathered in their patient's daily lives.

For instance, the data collected by InCense during the sensing campaign 1 does not encompass a thorough functional test about the patient; instead, it provides some glimpses of patient behavior, which can provide a different lens whereby physicians can follow or observe their patients in real-life conditions as their daily activities unfold. These data could be added to the existing healthcare ecosystem by sharing some information of interest to patients and their families, insurers, healthcare providers, and other stakeholders. In particular, some of these data can be of actual medical significance and therefore can be aggregated into the electronic medical record of the patient itself.

As is, the type of data collected, and presented in this chapter, can offer some insights into what the future may hold for the doctor's office, where decision-making processes could be a lot easier with reliable information.

5.6 CONCLUSIONS AND FUTURE WORK

Advances in mobile sensing are making possible the monitoring of activities and behaviors of large populations of mobile users. This information can be used to better understand how behavior affects health and to encourage individuals to adopt healthier habits. This requires advances in sensing hardware, pattern recognition algorithms, and the development of robust, flexible, and easy-to-use sensing platforms. As described in this chapter, however, performance issues still limit the applicability of mobile

sensing. Notably the battery and processing and storage limitations of mobile phones might prevent conducting certain mobile sensing campaigns.

We presented an extension to InCense, a research kit for collecting behavioral data from populations of mobile phone users. This extension was mainly aimed at supporting distributed sensing campaigns in order to optimize resources such as battery, storage, and bandwidth. In energy-craving environments, like mobile phone sensing, some of these resources are valued, and thus they pose serious challenges in sensing campaigns wherein significant amounts of data are to be collected (i.e., audio or video) or the battery is rapidly depleted (i.e., GPS). Given the limitations of mobile phones in these aspects, we implemented several strategies that helped us save processing time, storage, data transmission, and ultimately battery, thus making our sensing campaigns feasible.

Through the description of two sensing campaigns conduced with the use of InCense, we have illustrated how mobile sensing can be used to infer behaviors associated with the frailty syndrome among older adults and detect problematic behaviors in people suffering from dementia.

REFERENCES

[1] Larson, E.C., et al. Accurate and privacy preserving cough sensing using a low-cost microphone, in Proceedings of the 13th International Conference on Ubiquitous Computing (Ubicomp 2011), Beijing, September 17–21, 2011. ACM, New York, p. 375–384.

[2] Lu, H., et al., StressSense: detecting stress in unconstrained acoustic environments using smartphones, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Ubicomp 2012), Pittsburgh, PA, September 5–8, 2012. ACM, New York. p. 351–360.

[3] Dong, W., B. Lepri, and A. Pentland, Modeling the co-evolution of behaviors and social relationships using mobile phone data, in Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM 2011), Beijing, China, December 7–9, 2011. ACM, New York. p. 134–143.

[4] Beltran, J., et al., Detecting Disruptive Vocalizations for Ambient Assisted Interventions for Dementia, in Ambient Assisted Living and Daily Activities, L. Pecchia, et al., Editors. 2014, Springer International Publishing, New York. p. 356–363.

[5] Zavala-Ibarra, I. and J. Favela. Assessing muscle disease related to aging using ambient videogames, in Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2012), San Diego, CA, May 21–24, 2012. IEEE, San Diego, CA, p. 187–190.

[6] Tampi, R.R., et al., Behavioral and psychological symptoms of dementia: part II—treatment. Clinical Geriatrics, 2011. 19(6): p. 2–10.

[7] Sallis, J.F., N. Owen, and M.J. Fotheringham, Behavioral epidemiology: a systematic framework to classify phases of research on health promotion and disease prevention. Annals of Behavioral Medicine, 2000. 22(4): p. 294–298.

[8] Madan, A., et al., Social sensing for epidemiological behavior change, in Proceedings of the 12th ACM International Conference on Ubiquitous Computing (Ubicomp 2010), Copenhagen, Denmark, October 3–6, 2010, ACM, New York, p. 291–300.

[9] Dey, A.K., Understanding and using context. Personal and Ubiquitous Computing, 2001. 5(1): p. 4–7.

[10] Reddy, S., et al., Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks, 2010. 6(2): p. 1–27.

[11] Lane, N.D., et al., Enabling large-scale human activity inference on smartphones using community similarity networks (CSN), in Proceedings of the 13th ACM International Conference on Ubiquitous Computing (UbiComp 2011), Beijing, China, September 17–21, 2011. ACM, New York, p. 355–364.

[12] Sanchez-Cortes, D., et al., A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Transactions on Multimedia, 2012. 14(3): p. 816–832.

[13] Castro, L.A., et al., Behavioral data gathering for assessing functional status and health in older adults using mobile phones. Personal and Ubiquitous Computing, 2015. 19(2): p. 379–391.

[14] Clegg, A., et al., Frailty in elderly people. The Lancet, 2013. 381(9868): p. 752–762.

[15] Katz, S., et al., Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. JAMA, the Journal of the American Medical Association, 1963. 185(12): p. 914–919.

[16] Ware, J.E. and C.D. Sherbourne, The MOS 36-item short-form health survey (SF-36): conceptual framework and item selection. Medical Care, 1992. 30: p. 473–483.

[17] Rubenstein, L.Z., et al., Screening for undernutrition in geriatric practice: developing the short-form mini nutritional assessment (MNA-SF). Journal of Gerontology, 2001. 56A: p. M366–M377.

[18] Hallal, P., L.F. Gómez, and D. Parra, Lecciones aprendidas después de 10 Años del uso de IPAQ en Brasil y Colombia. Journal of Physical Activity and Health, 2007. 7(Suppl 2): p. S259–S264.

[19] Roper, W.L., et al., Strengthening the public health system. Public Health Reports, 1992. 107: p. 609–615.

[20] Fried, L.P., et al., Frailty in older adults evidence for a phenotype. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 2001. 56(3): p. M146–M157.

[21] Xue, Q.L., et al., Life-space constriction, development of frailty, and the competing risk of mortality: the Women's Health and Aging Study I. American Journal of Epidemiology, 2008. 167: p. 240–248.

[22] Schenk, A.K., et al., Cellular telephones measure activity and lifespace in community-dwelling adults: proof of principle. Journal of American Geriatrics Society, 2011. 59(2): p. 345–352.

[23] Peel, C., et al., Assessing mobility in older adults: the UAB Study of Aging Life-Space Assessment. Physical Therapy, 2005. 85(10): p. 1008–1019.

[24] Huang, X., J. Baker, and R. Reddy, A historical perspective of speech recognition. Communications of the ACM, 2014. 57(1): p. 94–103.

[25] Potamitis, I. and T. Ganchev, Generalized Recognition of Sound Events: Approaches and Applications, in Multimedia Services in Intelligent Environments, G.A. Tsihrintzis and L.C. Jain, Editors. 2008, Springer, Berlin, Heidelberg. p. 41–79.

[26] Krishnamurthy, N. and J. Hansen, Babble noise: modeling analysis and applications. IEEE Transactions on Audio, Speech and Language Processing, 2009. 17(7): p. 1394–1407.

[27] Beltrán, J. Activity recognition using a spectral entropy signature, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Ubicomp 2012), Pittsburgh, PA, September 5–8, 2012. ACM, New York, p. 576–579.

[28] Ibarrola, A.C. and E. Chavez, A robust entropy-based audio-fingerprint, in IEEE International Conference on Multimedia and Expo (ICME 2006), Toronto, Canada, July 9–12, 2006, IEEE, Toronto, ON, p. 1729–1731.

[29] Mitrović, D., M. Zeppelzauer, and C. Breiteneder, Features for Content-Based Audio Retrieval, in Advances in Computers, V.Z. Marvin, Editor. 2010, Elsevier, Amsterdam. p. 71–150.

[30] Beltran-Marquez, J., E. Chavez, and J. Favela, Environmental Sound Recognition by Measuring Significant Changes in the Spectral Entropy, in Pattern Recognition (LNCS 7329), J.A. Carrasco-Ochoa, et al., Editors. 2012. Springer, Berlin, Heidelberg. p. 334–343.

[31] Kim, D.H., et al., SensLoc: sensing everyday places and paths using less energy, in Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys 2010), Zurich, Switzerland, November 3–5, 2010, ACM, New York. p. 43–56.

6

PERVASIVE COMPUTING TO SUPPORT INDIVIDUALS WITH COGNITIVE DISABILITIES

Monica Tentori[1], José Mercado[1], Franceli L. Cibrian[1], and Lizbeth Escobedo[2]

[1] *Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico*
[2] *National Institute of Technology of Mexico, Mexico City, Mexico*

6.1 INTRODUCTION

Today is not a good day for Marley—a 5-year-old low-functioning child with autism[1] learning how to discriminate household objects. His teacher Bella has tried 10 times to teach him how to discriminate a glass, but Marley's cognitive impairments do not enable him to concentrate during therapies and meet the goals for the day.

> Bella starts the trial and places a glass and a fork on the table in front of Marley. Then, Bella starts the first of 10 trials asking Marley to grab the glass. Marley shakes his hands and moves his head from side to side looking around the classroom instead of looking at the objects. During this time Marley is "off-task." Bella physically re-directs Marley's attention to the therapy directing his head towards the objects, and prompting Marley saying, "Marley! Grab the glass!" while pointing towards the glass. Marley fidgeting, grabs the fork instead. Bella grabs Marley's hand and places it on the glass saying: "Marley! Grab the glass!." Marley grabs the glass and gives it to Bella. Bella rewards

[1] For simplicity of reading, the term "autism" will be used throughout this chapter to denote individuals within the spectrum [1].

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

him by saying: "Good job! Marley!" and gives him a piece of cookie. Then, Bella marks the first trial annotating a sad face in her notes to mark the first trial as incomplete, as Marley needed many prompts. When Marley sees the sad face on his notebook he gets angry and starts screaming at Bella.[2]

This is one of the numerous examples of the challenges individuals with cognitive disabilities and caregivers face in their everyday lives. The scenario also shows the amount of time and effort caregivers invest in helping individuals with cognitive disabilities reach their goals and monitor progress.

The World Health Organization (WHO)[3] indicates that 1–4% of the world population may have some level of cognitive disability, but this is difficult to confirm due to extreme variations in definition, culture, measurement sources, and methodology (Table 6.1). An individual with cognitive disabilities experiences the presence of incomplete or arrested mental development, principally characterized by the deterioration of concrete functions at each stage of development, and that contributes to the overall level of intelligence. Individuals with cognitive disabilities experience deficiencies in motor functioning, memory, executive functions, and the speed of cognitive processing such as learning, thinking, the use of language, and other mental functions. Cognitive disabilities are classified by functional or clinical disability.[4] Clinical diagnoses may be useful from a medical perspective for the treatment, but functional disabilities ignore the medical or behavioral causes of the disability and instead focus on the resulting abilities and challenges [2].

Developmental impairments and losing capabilities cause a decline in people's perceived quality of life affecting one or more of the basic self-care skills and well-being [3]. These deficits can in turn lead to lower perceived quality of life and self-care abilities. Indeed, individuals with cognitive disabilities often require constant assistance from caregivers[5] to help them successfully perform the activities of daily living (ADLs).[6] For example, one study indicated children with autism require help from a caregiver from a nearby distance of up to 3 ft for about 86% of the time [4–6], and the lack of independence also extends to other populations (e.g., elders with

[2] This scenario is a typical day of children with autism attending to Pasitos—a school clinic specialized in the care of children with autism located in Tijuana, Mexico.

[3] A specialized agency of the United Nations, established in 1948 and headquartered in Geneva, whose mission is to prevent the international spread of diseases, such as cholera, malaria, and poliomyelitis.

[4] Clinical diagnoses of cognitive disabilities include autism, Down syndrome, traumatic brain injury (TBI), and even dementia. Less severe cognitive conditions include attention-deficit disorder (ADD), dyslexia, and learning disabilities in general.

[5] For simplicity of reading, we will regularly refer to the stakeholders involved in the care of individuals with cognitive disabilities including relatives, therapists, psychologists, clinicians, and teachers as caregivers. Note that the literature distinguishes between formal and informal caregivers. Formal caregivers are those providers associated with a formal service system, whether a paid worker or a volunteer. Family or informal caregivers are any relative, partner, friend, or neighbor who has a significant personal relationship with, and provides a broad range of assistance for, the individual with cognitive disabilities.

[6] Activities of daily living (ADLs) is a term used in health care to refer to daily self-care activities within an individual's place of residence, in an outdoor environment, or both. ADLs include activities such as bathing, eating, sleeping, dressing, grooming, work, homemaking, and leisure.

TABLE 6.1 Most Common and Prevalent Types of Disabilities Affecting Cognitive and Intellectual Functioning: Definition, Symptoms, and Impairments

Definition	Impairments and Symptoms
Dyslexia is the most common form of language-based learning disability. It is primarily a reading disability, and there is evidence suggesting that the condition is inherited.	Difficulty in single world decoding, spelling, and writing. Problems with language and reading.
Attention Deficit Hyperactivity Disorder (ADHD) is a medical condition affecting a person's ability to focus, sit still, and pay attention.	Difficulty in focusing on tasks or subjects and following instructions or assigned tasks. Fidgeting, feeling restless, or talking excessively.
Brain injury is any injury occurring in the brain of a living organism including the following: • *Brain damage*, the destruction of degeneration of brain cells; • *Traumatic brain injury*, damage that occurs when an outside force traumatically injures the brain; • *Stroke*, a vascular event causing damage in the brain; • *Acquired*, damage to the brain occurred after birth.	Each brain injury is unique so the extent of the injury to the person's brain determines the outcome of the person's ability to process information. Common impairments include the following: • Physical problems; • Swallowing, sensory, communication, and cognitive deficits; • Behavioral changes.
Autism is a neurodevelopmental disorder characterized by impaired social interaction, verbal and nonverbal communication, and restricted and repetitive behavior. It is usually accompanied with other disorders such as sensory processing disorders, impairments in motor functioning, and attention deficit disorder.	Depends on the level of functioning: low, medium, or high; common impairments include the following: • Lack of eye contact; • Speech and language problems; • Difficulties in social interactions; • Stereotypy, ritualistic, and restricted behavior; • Self-injury; • Sameness and fidgeting.
Alzheimer's disease is a chronic neurodegenerative disease in which the death of brain cells causes memory loss and cognitive decline. A neurodegenerative type of dementia, starting mild and getting progressively worse.	Depending on the stage: early, moderate, or severe; common impairments include the following: • Short-term memory loss; • Problems with language; • Disorientation • Mood swings; • Behavior issues.

Alzheimer's disease or AD [7, 8]). Caregivers following individuals with cognitive disabilities to provide directions, prompts, and guidelines help them accomplish their goals, but this type of support might, paradoxically, hamper their independence and opportunities to fit into the society. Indeed, the ability to accomplish the ADLs without physical help is considered a requirement for independent functioning and self-management [9].

To support individuals with cognitive disabilities, researchers have explored a wide range of pharmacological and nonpharmacological interventions. Even though nowadays there is no cure for cognitive impairments, early intervention may help some individuals to integrate into the society and increase their chances to live a typical life. In particular, nonpharmacological interventions have been widely focused on three major areas:

- Supporting caregivers when monitoring parameters relevant for clinical case assessment [10–12];
- Understanding how technology could promote the independence of individuals with cognitive disabilities [4, 5] and augment innovative therapies [13, 14];
- Facilitating the rehabilitation and development of motor skills [15–17].

First, some efforts have been devoted to developing innovative methodologies and tools to facilitate the measurement of behaviors that are relevant to clinical case assessment. Many individuals with cognitive or developmental disabilities exhibit numerous behavior problems. For example, children with autism often exhibit maladaptive behaviors that could be internalized (i.e., emotionally reactive, depressed/anxious affect, somatic complaints, and withdrawal) or externalized (i.e., aggression, defiance, and inattentive) [18], and patients with AD or dementia also exhibit agitation, aggression, mood disorders/behavioral disturbance, apathy, depression, psychosis, and hallucinations [19]. Exhibiting these behavior problems increases the risk of harm or injury to both the individual and those nearby. Many treatment programs have been developed to reduce the frequency and severity of these problems, demanding the collection of accurate data annotating with as much level of detail the frequency and severity of the observed behavior to understand why and when it occurred, or to determine if there is a change in the behavior as a result of treatment [20]. The most common methods for measuring behavior problems rely on self- or caregiver-provided status reports, direct observations, or brief physical tests that are potentially biased, infrequently obtained, and cannot capture detailed patterns of movements or activities. Alternatively, other approaches make use of video recording to solve some of the problems of missing data during direct observation opening the possibility to consult the recordings later and assess the behavior of the individual. However, the continuous recording of videos generates several problems including limited amount of storage and the need for too much time and effort to browse through irrelevant video material. New ways to measure relevant behaviors and intelligently segment videos are essential to evaluate future treatments for individuals with cognitive disabilities and collect empirical evidence of detailed behaviors.

Other projects have explored how to augment cognitive behavioral therapy (CBT)—a psychotherapeutic approach that addresses maladaptive cognitive processes through a number of goal-oriented and systematic processes making use of visual supports and promoting repetition [21, 22]. Visual supports [23] typically include the exchange or display of a variety of images, drawings, or photographs on laminated cards helping individuals with cognitive disabilities to communicate needs and giving them structure by representing tasks and goals that must be followed in an order. Technological interventions that provide interactive visual supports (e.g., vSked [24] and MOSOCO [25]) including the annotation of text on top of physical objects [26], the use of verbal and physical prompts, and rewards help caregivers to keep individuals with cognitive disabilities "on task" [24]. However, these tools still lack some of the realism real objects have to help children with autism generalize from the classroom to other environments. Thus, a new type of interactive visual support capable of augmenting the physical form of a traditional object with digital and interactive information is needed to combine the benefits of both digital and physical supports.

Others have explored how to develop interactive systems to help individuals practice gross motor coordination skills that rely on postural control and mobility [27]. The lack of motor coordination may severely limit the mobility, navigation, and the emergence and performance of other motor skills. Movement enables individuals to satisfy basic needs, communicate, and conduct ADLs. Motor skills are commonly used as a quality-of-life measurement, as they are critical for an individual's independence [28, 29]. Current available interventions to support motor impairments for individuals with cognitive disabilities are either process driven or goal oriented.[7] Process-driven motor interventions support the development and acquisition of sensory, perceptive, integration, attention, and memory functions involving sensory integration therapy [30], perceptual motor [31, 32], and kinesthetic stimulation [33]. In contrast, goal-oriented interventions focus on the development of specific skills (e.g., jump, walk, and run) by teaching or promoting the practicing of basic motor skills through repetition [34–36]. But repetition is boring for individuals with cognitive disabilities who exhibit poor adherence to motor therapeutic interventions. Exergames[8] are appropriate to help neurotypical[9] individuals practice motor skills [37], especially because they are engaging [38]. However, available commercial exergames and those reported in the literature might not be appropriately designed to be used by individuals with cognitive disabilities. Most available exergames present a complex, dynamic gameplay, difficult to be understood by individuals with cognitive disabilities and excessive cognitive load, as available exergames have multiple stimuli and features. Designing exergames to meet the needs of individuals with cognitive disabilities is not an easy task [39].

[7] A therapy oriented to goals or goal-oriented involves detailed planning of what the person will do, including a definition of the behavior specifying frequency, intensity, or duration.

[8] The activity of playing video games that involve physical exertion and is thought of as a form of an exercise.

[9] An individual who does not have an atypical neurology. In other words, anyone who does not have conditions such as autism, dyslexia, development coordination disorder, ADHD, etc.

Undeniable, traditional technology has played a significant role in supporting nonpharmacological interventions, but it is clear that a new type of technology is available "anywhere and anytime," [40] enabling the integration of the digital and the physical world to support cognition, improving engagement, and encouraging repetition during motor rehabilitation therapies; innovative ways to measure data relevant for clinical case assessment are needed.

Mobile, pervasive, and ubiquitous computing (Ubicomp) technologies offer promising solutions to documenting progress, diagnosing conditions, and treating and managing care-related activities. Pervasive computing, ubiquitous computing, and ambient intelligence are concepts evolving from the development and deployment of pervasive applications, frequently in the healthcare domain and are most of the time mentioned in the healthcare context [41]. Initial visions of pervasive and Ubicomp describe environments furnished with computational artifacts that remain in the background and have intelligent capabilities to support user-centered activities [42–45]. Pervasive health [46][10] applications range from wearable and embedded sensors that assist in self-care [47–52] to hospital environments enhanced with pervasive computing technology [53–57].

In this chapter, we analyze a variety of research projects and commercial solutions devoted to the understanding, designing, developing, and pilot testing pervasive healthcare applications to support individuals with cognitive disabilities. Our review included an exhaustive search in relevant research libraries including the ACM digital library[11] and the revision of papers published in top-tier conferences, magazines, and journals, in Ubicomp,[12] human–computer interaction (HCI),[13] and health care, particularly those forums with special emphasis in supporting individuals with cognitive disabilities and impairments. Our aim, here, is not to provide a comprehensive review of all of the relevant literature in this vast interdisciplinary space. Instead, this chapter describes the field of pervasive computing to support individuals with cognitive disabilities in a way that prepares newcomers to tackle some of the most important issues and identify upcoming trends in this important research space. The general inclusion criteria for the projects discussed in this chapter included papers discussing issues related to the following:

- Mobile devices (e.g., laptops, PDAs, tablet PCs, and mobile phones) and cognitive assistants supporting a challenge associated to cognition;
- Wearable and on-body sensors (computer-enhanced textiles or medical sensors) to support the monitoring of data;

[10] Pervasive health care is defined as the "use of mobile/wearable/environmental technologies that have come out of the ubicomp/pervasive research communities that is targeted at some challenge associated to health. The health part of it could be addressing issues of diagnosis or treatment that leads to increased medical knowledge (what we would consider medical or biomedical research) or it could be related to work that is contributing to the science or engineering of healthcare delivery."

[11] ACM digital library http://dl.acm.org/.

[12] Ubiquitous computing (Ubicomp) is a concept in software engineering and computer science where computing is made to appear everywhere and anywhere, in a way that is "invisible" to the user.

[13] Human–computer interaction (HCI) researches the design and use of computer technology, focusing particularly on the interfaces between users and computers.

- Exergames to support rehabilitation, exertion, and motor skills development;
- Pilot studies and case studies conducted in healthcare settings for understanding healthcare needs and the use of pervasive applications of individuals with cognitive disabilities.

First, we will describe how wearable and mobile sensing platforms could ease the manual record-keeping of behavioral data [58–63], moods, and emotions [64–66] and enable clinicians to sense the body [20, 59, 62, 67, 68] uncovering the data relevant for clinical case assessment. Second, we will describe solutions integrating the physical and the digital world, like augmented reality (AR)[14] and tangible computing[15]; in addition, the use of mobile applications to support cognitive therapies [25, 71–73] provides step-by-step guidance [74–76] and helps nonverbal individuals with alternative solutions to augment communication [77–82]. Next, we will describe solutions leveraging the use of exergames and serious games to support engagement [83–85] and the development of age-appropriate motor skills [86–89] and maintain the adherence of individuals with cognitive disabilities during motor and rehabilitation therapies [90–93]. Last, we will describe several projects working toward helping disoriented individuals to navigate in both indoor [94–98] and outdoor environments [99–102].

To further detail the potential benefits of using pervasive technology in support of individuals with cognitive disabilities, we present three case studies showcasing how pervasive computing could augment traditional nonpharmacological therapies and improve some of the challenges faced by individuals with autism and elders with AD. The first case study describes how elders with AD living in a nursing home make use of a sensing toolkit running in mobile phones to facilitate the gathering of behavioral data that could be later used by geriatrics for clinical case assessment [103]. The second case study shows how mobile AR technology could help children with autism reduce their behavioral problems and increase attention during cognitive therapies [104]. Finally, the third case study describes how an exergame helps individuals with autism to be engaged when practicing upper-limb movements to support their motor skills development [105].

The information presented in this chapter is based on the analysis of the literature as well as our own research experiences and those of many of our colleagues. We will mainly focus in discussing projects related to support children with autism and older adults with AD, as these populations are considered the fastest growing cognitive disabilities in the United States—5.1 million[16] of Americans have AD and 3.5 million[17] Americans live with autism spectrum disorder (ASD). Also pervasive healthcare research has been more active in support of these populations.

[14] Augmented reality (AR) allows the user to see the real world augmented with virtual objects superimposed upon or composited on top of the real world [69].

[15] Tangible computing proposes that a computer is embedded into objects offering natural physical world interactions [70]. Tangible computing makes use of a tangible user interface in which an individual interacts with digital information through the physical environment.

[16] Available at http://www.alzfdn.org. Date accessed: March 31, 2015.

[17] Available at www.autismsociety.org. Date accessed: March 31, 2015.

6.2 WEARABLE AND MOBILE SENSING PLATFORMS TO EASE THE RECORDING OF DATA RELEVANT TO CLINICAL CASE ASSESSMENT

Body motion tracking techniques mainly involve the use of "invasive" sensors attached to our body [106, 107] or three-dimensional (3D) cameras that noninvasively [76, 108, 109], either reading special tag markers [76] or reconstructing individuals' body [108], track our body and movements. Most of this body of work has been focused on identifying physical exercises [110], compensation when exercising [111], body postures [68, 108], body expressions [112], and different behaviors [20, 59] and activities [62, 76, 109]. This body of work has also explored the use of capture and access tools and mobile sensing platforms to ease the manual record-keeping of behavioral data [58–63], mood, and emotions [64–66] and enable clinicians to sense the body [20, 59, 62, 67, 68] uncovering data relevant for clinical case assessment.

6.2.1 Capture and Access Tools

Different projects have explored how pervasive computing environments help clinicians to capture behaviors and manage video recordings by selectively specifying annotations or landmarks to extract relevant data from video recordings and sensor data.

For example, ABARIS [61] is an application created to support clinicians when capturing behaviors children with autism exhibit during applied behavior analysis (ABA)[18] therapies. ABARIS is a system using a webcam that captures video and audio data; a wireless microphone is used to capture commands and mark the beginning of sections during therapy using voice recognition techniques, and a digital pen captures annotations the therapist records during the session. All the captured information is stored in a computer and can be accessed by several specialists individually or collectively when discussing the progress of a patient. The use of the system showed a system like ABARIS improves the practices of therapists, allowing better attention to the therapy itself, since the use of the system reduces the time therapists invest with record-keeping. Altogether, ABARIS allows better assessments and group discussions on some situations. Similarly, the CareLog system was designed and implemented to be a pervasive video monitoring technology capturing four angles of a classroom and automatically generating graphs, showing when and how often a particular behavior occurs; CareLog supports the functional behavioral assessment (FBA).[19] The evaluation of CareLog shows the system is usable, teachers

[18] Applied behavior analysis (ABA), previously known as behavior modification, is an applied natural science devoted to developing and applying procedures for effective and beneficial behavior change. It involves the use of empirically demonstrated behavior change techniques to increase or decrease the frequency of behaviors, such as altering an individual's behaviors and reactions to stimuli through positive and negative reinforcement of adaptive behavior and/or the reduction of behavior through its extinctions, punishment, and/or satiation [113].

[19] Functional behavioral assessment (FBA) is a practice in which caregivers work to try to understand the function of inappropriate behaviors [114], an attempt to look beyond the obvious interpretation of behavior as "bad" and determine what function it may be serving for a child.

needed minimal training to use the system, and it positively impacted teachers' workload [10]. Despite the potential danger of being an invasive application, the system resulted in minimal intrusion to the environment and the teaching activity with substantial benefit to the documentation, monitoring, and diagnosing of behaviors.

Undeniably, the design of capture and access tools for monitoring behaviors and parameters relevant to clinical case assessment face several challenges. First, it is not clear how to identify which information is worth capturing and how these moments of interest could be automatically detected. Reviewing videos captured 24 h a day, 7 days a week is a time-consuming task requiring an enormous amount of storage capacity. Therefore, we must develop mechanisms that allow the automatic recognition of relevant events to be archived and indexed, or novel tools to facilitate the manual record-keeping. Second, direct observation requires extensive training and reliability assessments, and, unfortunately, observed behaviors are still very subjective and somewhat arbitrary. Some behaviors can be especially difficult to recognize based on what they look like, while others are difficult to track accurately and objectively. There is no way to objectively assess the intensity of a behavior by human observation alone [20].

6.2.2 Sensing the Body and the Environment

All these drawbacks have been the motivation in recent decades for a significant amount of proposals of new tools and techniques to more accurately monitor and infer behavior problems and activities, which in turn is the key to planning and evaluating therapies or treatments and identifying and diagnosing individuals with disabilities.

This body of work combines the use of body sensors (e.g., accelerometer, gyroscope, electrocardiography (ECG), and galvanic skin response (GSR)) and ambient sensors (e.g., camera, microphone, pressure sensor, and infrared motion sensor) for data acquisition. All these sensors are usually connected to smartphones, tablets, and personal computers responsible for data storage and to provide the adequate visualizations of data to the end user (see Case Study 6.1). Some of these solutions make use of complex pattern recognition and prediction algorithms[20] (e.g., voice recognition [115], support vector machine [116], decision trees [117], and hidden Markov model (HMM) [118]) to infer more complex representation of behavioral data [20, 59, 119]. For example, available solutions for activity recognition have largely been focused on the automatic recognition of ADLs—like handwashing [109] or meal preparation [120]. Lee and Mase reported a dead-reckoning method to recognize and classify a user's sitting, standing, and walking activities using the acceleration and angular velocity of wearable sensors on the users' body [62].

Similarly, other projects have used wearable and implantable body sensors to acquire health data, which in combination with wireless networks and consuming

[20] It refers to a branch of machine learning that focuses on the recognition of patterns and regularities in data, although it is in some cases considered to be nearly synonymous with machine learning.

TABLE 6.2 Several Body Sensors Measuring Physiological Signs

Sensor	Measurement
Accelerometer	Direction
Gyroscope	Orientation
Image/video	Activity
Glucometer	Blood sugar
pH meter	pH of a liquid
CO_2 gas sensor	CO_2 concentration
Electrocardiography (ECG)	Cardiac activity
Electroencephalography (EEG)	Brain activity
Electromyography (EMG)	Muscle activity
Electrooculography (EOG)	Eye movement
Pulse oximetry	Blood oxygen saturation
Galvanic skin response (GSR)	Perspiration
Thermal	Body temperature

Table modified from Ref. [121].

electronic technologies (e.g., smartphones, tablets, laptops, and personal computers) have the potential to change the lifestyle of individuals with cognitive disabilities (Table 6.2, Ref. [107]). Body sensors can capture a lot of contextual information that when used in tandem with alerting mechanisms of abnormal conditions could support their independence, following notions like aging in place or promoting the practicing of therapies at home. Also, the contextual information obtained from these sensors could support decision making when clinicians and caregivers decide how to change a treatment or emit a diagnostic. For example, the work of Plöetz *et al*. [20] describes a technique for using on-body accelerometers to assist in the automated classification of problem behavior during an intervention using direct observation with children with developmental disabilities. They use machine-learning techniques to segment relevant behavioral episodes from a continuous sensor stream and to classify them into distinct categories of severe behavior (aggression, disruption, and self-injury). Their results show promising classification results when the sensing and analysis system is applied to data from a real assessment session conducted with a child exhibiting behavior problems.

Overall, the literature has proposed two ways for monitoring health metrics using body sensors. First, episodic monitoring refers to the recording of specific indicators to track the progress of a disease or recovery. These technologies involve tracking vital signs and the presence of specific indicators. For example, Goodwin *et al.* [67] evaluated the use of wireless three-axis accelerometers and pattern recognition algorithms to automatically detect body rocking and hand flapping in children with ASD. This could enable researchers and clinicians to systematically study what functional relations exist between these behaviors and specific antecedents and consequences. These measurements could also facilitate efficacy studies of behavioral and pharmacological interventions intended to replace or decrease the incidence or severity of stereotypical motor movements.

Second, other projects have promoted the continuous monitoring for people who need frequent or constant monitoring. These technologies involved wearable and implantable body sensors that increase the early detection of emergency conditions and diseases and also provide healthcare services for people with various degrees of cognitive and physical disabilities. For example, Lai *et al.* [68] used several triaxial acceleration sensor devices for joint sensing of injured body parts for elderly falling. The proposed model can determine the possible occurrence of fall accidents. In addition, after a fall accident occurs, the impact acceleration and normal acceleration can be compared to determine the level of injury. They also implemented a sensing system for analysis where the areas of the body parts that may sustain greater impact force are marked red so that more information can be provided to medical personnel for more accurate judgment and for reference in giving first aid.

All in all, pervasive healthcare applications for body sensors involve providing interfaces for the disabled, integrated patient monitoring, diagnosis, drug administration in hospitals, telemonitoring of human physiological data, and tracking and monitoring patients and/or doctors inside homes [106].

In deciding the type of sensors to use in any type of pervasive healthcare system, it is important to understand the user population and its requirements. Some individuals with cognitive disabilities experience sensory processing disorders and will not tolerate to wear sensors attached to their body. Additionally, fidgeting and stereotyped movements could add error to the sensed data, affecting recognition accuracy and precision. Open questions remain to understand what is the appropriate hardware for this population, and the development of novel approaches for noise reduction when working with individuals with cognitive disabilities exhibiting "atypical" movements. Sometimes, the use of wearable and mobile platforms needs to be unobtrusive as much as possible and embedded into the environment. An alternative for this is the use of computer vision techniques that are capable to avoid any further cognitive load on the user by requiring some input or effort [76], but may impose privacy concerns as a downside.

Computer vision techniques are also used for behavioral assessment [108] and to monitor activities [76]. These techniques also allow the representation of more accurate models of interaction with the environment. Nevertheless, issues surrounding the use of vision need to be addressed in order for such systems to become more widely accepted and used. For example, Hashemi *et al.* [108] used a computer vision approach for the assessment of autism-related behavioral markers[21] based on components of the autism observation scale for infants (AOSI)[22] [122]. They develop algorithms to measure three critical AOSI activities that assess visual attention and augment these activities with an additional test that analyzes asymmetrical patterns

[21] Behavior patterns typically exhibited by individuals with autism. These include atypicalities or delays in social communication behaviours (e.g., anticipatory social response, social babbling, orientation to name, and eye contact) and nonsocial behaviors (e.g., disengagement of visual attention, motor control and behavior, and atypical sensory behaviors) as well as aspects of temperament (e.g., reactivity and ease of transitions between activities).

[22] It is a semistructured, experimenter-led behavioral assessment designed to measure early behavioral markers of ASD in infants aged between 6 and 18 months.

in unsupported gait. The algorithms involve assessing head motion by facial feature tracking, and the gait analysis relies on joint foreground segmentation and 2D body pose estimation in video. They show results that provide insightful knowledge to augment the clinician's behavioral observations obtained from real in-clinic assessment.

These include technical issues such as the development and optimization of image processing algorithms, as well as ethical and social issues such as those concerned with the privacy of users and the use of vision in the home or clinics where vision-based systems support the monitoring of behaviors and the ADL.

6.2.3 Leveraging the Capturing of Moods and Emotions

Emotions play a significant role in the expression of human intelligence and represent a challenge to individuals with disabilities that due their cognitive decline have problems to recognize and express emotions. The arguments in support of the significance of emotions introduced a new area of "emotional intelligence," defined as the capacity to understand emotions and to reason with emotions [123].

Emotional intelligence is covered by the field of affective computing, which represents "computing that relates to, arises from, or deliberately influences emotions" [65]. In other words, it is trying to assign to computers human-like capabilities of observation, interpretation, and generation of affect features. Affective computing is an interdisciplinary field, spanning computer science, psychology, and cognitive science [124]. The origins of the field may be traced as far back as to early philosophical enquiries into emotion, but the more modern branch of computer science originates with Rosalind Picard's 1995 paper "Affective Computing"[23] [65]. Affective computing has many applications to support individuals with cognitive disabilities—for example, computer-assisted learning, perceptual information retrieval, entertainment, and human health. Negative emotions like stress, frustration, irritation, and depression weaken the human immune system. Moreover, positive emotions highly contribute to physical and mental health. The awareness of the significance of emotions and their impact on human health has boosted the motivation for improvement of the research of this field [66, 125].

In this regard, pervasive computing literature has explored how to measure and infer emotions of individuals with cognitive disabilities and their caregivers. There is evidence showing that we can build systems that begin to identify both emotional expression and its generating state by measuring physiological responses that often arise when expressing emotions. Most widely acknowledged forms for identifying the emotional state are the analysis of facial expressions, voice, motor output, and data from affective wearable computers (e.g., heart rate, diastolic and systolic blood pressure, pulse, pupillary dilation, respiration, skin conductance, and temperature). In general, the input data are obtained from images, videos, audios, and some

[23] A more comprehensive introduction and explanation of affective computing can be found in the book of the same name published by Rosalind Picard [65].

sensors, and then using pattern recognition and machine learning techniques classification models a system identifies emotions in new input data.

All in all, there are three main methods for emotion research: *neurophysiological*, which measures signals such as the brain activity and pulse rate; *observer*, which measures facial expression, speech, and gestures; and *self-report* methods, which aim to describe emotions through interviews, questionnaires, and self-reported diaries [65, 126]. Technological support to record emotions may be done through typical capture and access tools into which a user inputs his/her moods through a graphical user interface (GUI) or other times using specialized devices (e.g., tangible user interfaces). For example, the Q Sensor from Affectiva[24] is a wearable biometric sensor in use at leading universities and corporations, which tracks user's excitement, engagement, stress, and anxiety by measuring slight electrical changes in the skin, known as electrodermal activity. The Q Sensor is worn in a wristband and lets people keep track of stress during everyday activities, giving doctors, caregivers, and patients themselves a new tool for observing reactions. The device also has a temperature sensor to help correct for mistakes, a clock, a rechargeable battery that lasts a day, an external button that lets a person put an event marker on the data, and a motion sensor that tracks movement in three directions. Caregiver or user can download data via USB or Bluetooth and can use a software to view, compare, and annotate the data with descriptions of events during high- and low-stress periods.

To illustrate an example of a mobile sensing platform that facilitates the capturing of events relevant for clinical case assessment, here we present the case of InCense. InCense helps the deployment of sensing campaigns leveraging the sensors available in smartphones to monitor physical activity and mobility data.

Case Study 6.1 A Mobile Sensing Toolkit Facilitating the Deployment of Sensing Campaigns in Mobile Phones

Leveraging data from previous user studies and conducting several multidisciplinary design sessions, members from our lab designed and developed InCense (Fig. 6.1, Ref. [103]). InCense is a toolkit facilitating the collection of behavioral data gathered by populations of mobile phone users. It was designed to help researchers with low technical skills and interested in conducting large-scale social and behavioral studies. It can handle numerous mobile participants without affecting the performance of the system.

InCense has a friendly user interface where researchers can drag and drop icons representing the sensors available in a mobile phone, like global positioning system (GPS), accelerometer, and Bluetooth. After placing the "icon-sensors" on the GUI, researchers may connect them to user actions or other sensors and define thresholds to activate data capturing. For example, activate the accelerometer when the GPS indicates the user enters a specific location. The collected raw data can be sent to a server running pattern

[24] http://www.affectiva.com/.

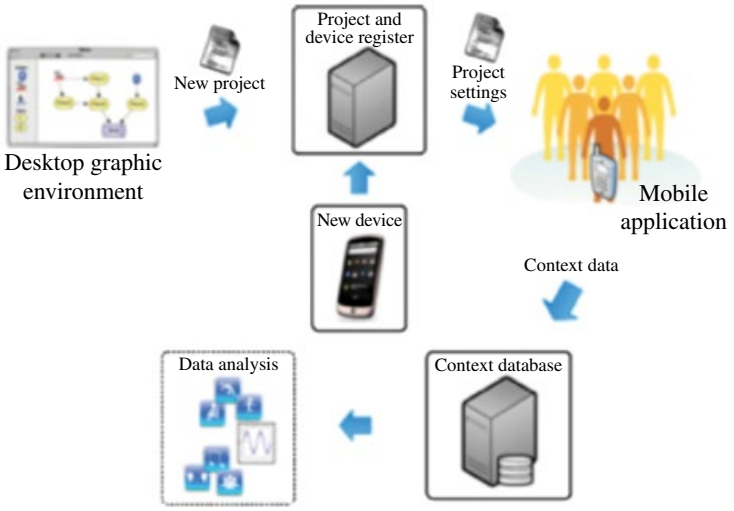


FIGURE 6.1 The high-level description of the InCense architecture (top) and older adults carrying InCense-enabled smartphones (bottom).

recognition algorithms to enable the inference of more complex contextual information, such as sleeping behavior or level of physical activity. InCense also enables researchers to create surveys that could be prompted based on sensor activity—for example, when a user's activity is low, ask "In the last week, how many days did you walk at least 10 min?"

To evaluate InCense, a sensing campaign was deployed in a nursing home of elders with AD, and the data gathered through the sensing campaign were

compared against traditional tools being used for clinical case assessment. Fifteen older adults participated during the evaluation of InCense. The participants' average age was 75.3 (SD = 1.8), ranging from 73 to 79 years. The sensing campaign was deployed to detect activity and mobility patterns that may indicate signs of wandering. Wandering is a common behavior among older adults with AD, and it represents one of the main concerns of caregivers because the person may get lost or injured. The frequency and manner in which an older adult wanders is highly influenced by the person's background and contextual factors specific to the situation. Wandering is an indicator of frailty. Traditional tools to measure wandering use the Timed-Up-and-Go (TUG) test—a simple test used to assess a person's mobility and requires both static and dynamic balance. The results indicate the data gathered with InCense were equally useful and helpful than the data gathered with the TUG test to support clinician's assessment and diagnosis/prognosis. In addition, clinicians explained that with InCense the granularity and periodicity of gathering daily data is better than the one obtained with the TUG test performed on a monthly basis—as individuals mostly travel to a specialized laboratory for data measurement based on their availability and mobility resources.

6.3 AUGMENTED REALITY AND MOBILE AND TANGIBLE COMPUTING TO SUPPORT COGNITION

Projects in pervasive health care to support cognition have proposed several solutions integrating the physical and the digital world, like AR and tangible computing; as well as the use of situated displays and mobile applications to support cognitive therapies [25, 71–73], provide step-by-step guidance [74–76], and help nonverbal individuals with alternative solutions to augment communication [77–82].

6.3.1 Prompting and Reminders to Provide Step-by-Step Guidance

Early intervention is crucial to promote independence and teach appropriate skills to support self-care management [127]. Self-care is one of the biggest challenges for individuals with cognitive disabilities [128]. The use of prompting and reminders in tandem with visual schedules is an appropriate tool to provide step-by-step guidance and help individuals with cognitive disabilities when conducting self-care activities. Visual schedules are a series of visual supports representing steps of activities or tasks. Each visual support in the schedule is a visual prompt, but caregivers also verbally prompt individuals with cognitive disabilities to provide reassurance when conducting activities. When a step is completed, the caregiver marks the task, or removes the visual support, to indicate which steps have already been completed and which steps remain. Although this strategy is effective for individuals with cognitive disabilities, it demands a lot of effort from caregivers who spend a considerable amount of time setting up the environment to attach the visual supports in the location where the individuals might need them. Step-by-step guidance and schedule management pervasive services could be useful to remind people how to adequately execute their activities.

FIGURE 6.2 PEAT [130] smart agenda (left), COACH [75] step-by-step guidance in intelligent environments (right).

Computer-based devices leveraging the use of visual schedules are being used to assist individuals with cognitive disabilities when performing ADLs [74–76, 109, 129]. For example, Levinson [130] designed and developed the Planning and Execution Assistant and Training system (PEAT) (Fig. 6.2 left) running in a mobile device to help individuals with traumatic brain injury (TBI) maintain their independence during the execution of ADLs. The PEAT system is a memory notebook that helps users to stay on task. To do this, the system uses a cue card that provides information about the current executed activity including people involved, date, and how to start and finish the activity. This system uses deterministic classical planning algorithms to calculate the best plan to complete a task and provide step-by-step guidance through visual and audible clues. Similarly, Cognitive Orthosis for Assisting aCtivities in the Home (COACH) [75] (Fig. 6.2 right) is designed to assist with the activity of hand washing to individuals with AD or severe dementia who have difficulty remembering the proper sequence of ADLs. The COACH system uses a video camera to observe the user as he/she attempts to wash his/her hands. The video image is processed to identify the step currently executed (e.g., turning on the water, using soap, and drying hands). If the system recognizes a problem—for instance, that the subject is using the towel before wetting his/her hands—a prerecorded verbal prompt is provided. Another version of the COACH system has been tested in support of children with autism [129]. The system uses a Markov decision to determine the activity and refine audible prompts and video demonstrations.

Overall technologies that provide step-by-step guidance have demonstrated the potential to teach self-care skills for individuals with cognitive disabilities and are appropriate to augment traditional visual schedules. This kind of support can also be useful for caregivers as a tool to teach crucial life skills while alleviating caregivers' burden and workload.

6.3.2 Language and Augmentative and Alternative Communication

Language and communication are two of the main problems of individuals with cognitive disabilities—mainly for children with autism as they have not yet developed language [131]. To this aim, mobile and pervasive computing technologies have explored how to support language and verbal communication to help individuals develop age-appropriate language and communication skills [81, 132]. Other projects have explored how to use augmented and alternative communication (AAC) devices [78–81, 133] to promote the practicing of language skills like improving pronunciation [134, 135].

There is an extensive literature [78–81, 133] in the AAC that addresses the communication needs of individuals with cognitive disabilities, mainly individuals with autism [77]. In some of these studies, children with autism use a device that pronounces words or short phrases on their behalf to communicate what they want or need [78–80]. For example, Dynavox [136] is a speech-generating device (SGD) using voice output helping individuals with autism to communicate their needs (Fig. 6.5). The use of SGDs shows these devices become a vital tool for individuals with cognitive disabilities and are appropriate to improve their communication skills [80]. This technology works very well in a classroom or at home where the SGD is connected and located in a stationary manner, but some users regret the lack of capabilities to be used in mobile devices. In this direction, Proloquo2Go (Fig. 6.3, Ref. [137]) and TapToTalk[25] are mobile AAC systems running in an iPhone or iPod touch to meet diverse communication needs of people with autism who struggle with speech skills. These two systems are a great support to this

FIGURE 6.3 A child with autism using Proloquo2Go by pressing buttons that are figures referring to a word or phrase the child wants to say and communicate [81].

[25] http://www.taptotalk.com.

community in mobile and dynamic contexts enabling them to communicate and interact socially.

Although all these devices have proved to be effective as an alternative way to support language and communication, there are open questions as to how individuals with cognitive disabilities using the AAC could generalize the newly learned skills or if they will be able to communicate at all without using or needing the AAC device. Such a device could create a high level of dependence, and it is not clear if this sort of "assisted communication" could help individuals communicate in a typical way, as the device will be in the way most of the time. Even though nowadays we have been somehow increasingly depending on our own mobile phones, this trend of building systems that will replace basic self-care activities with a device could create a world where technology will replace our actions making us less and less capable of satisfying our own needs.

6.3.3 Supporting Cognitive Training

Work on pervasive computing has explored the use of a variety of technologies (e.g., livescribe,[26] near-field communication (NFC), and radio-frequency identification (RFID) to augment visual supports in the form of "digital labels" (e.g., TouchCounters [138], Memory Spot [139], and Tap and Play [140]) by tagging physical objects with digital information to integrate the physical and the virtual world. A digital label acts as a "*bridge between the physical and virtual worlds [connecting] objects to services and applications. […] labels are realized through tags—physical entities attached to or integrated with objects* [30]." A recent trend in creating "digital labels" to integrate the physical and the digital world involves the use of AR to enable a user to use a smartphone as a "visor" to discover the digital information available in a physical object (e.g., sentient visors [16]). AR has been successfully used with children with autism (e.g., MOSOCO [25] and ARVe [141]); however, their efficacy is yet to be tested and open questions remain as to how AR could combine the benefits of both physical objects and interactive digital information to support real-time structured lessons for educational interventions.

Some lines of inquiry of our own research have been devoted to understanding the characteristics of visual supports and creating a new form of physical labels using AR and tangible computing through the development of MOBIS (see Case Study 6.2), and Things That Think (T3, Fig. 6.4, Ref. [71]). T3 is a 1.2″ cube embedding an accelerometer, a loudspeaker, a motor, a multicolor led array, and a microphone supporting the discrimination therapy of children with autism. T3 is a smart cube that converts any object we have created or someone else has crafted in an interactive toy that illuminates, moves, talks, and sings when appropriate. The results of the use of T3 by 18 children with autism for 2 months show smart objects reduce the workload of teachers, ease the record-keeping and increase its reliability, and reduce the behavioral problems of children with autism [71].

[26] http://www.livescribe.com/en-us/.

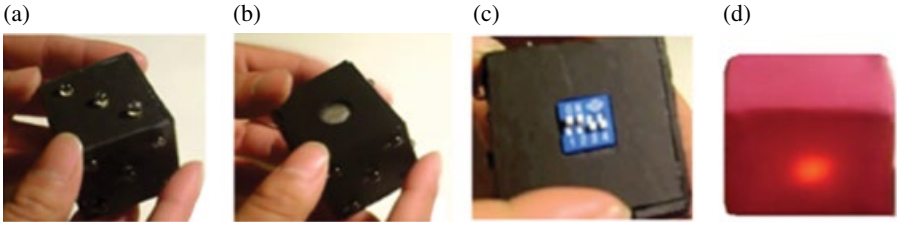


FIGURE 6.4 The T3 cube integrating a led array (a), a loudspeaker (b), and a configuration panel (c). A T3 object embedding a T3 cube blinking (d).

To illustrate an example of an AR technology used by individuals with cognitive disabilities and their benefits related to health, here we detail a case study exemplifying how an AR could help individuals with autism during cognitive training.

Case Study 6.2 A Mobile AR System to Support the Cognitive Training of Children with Autism

Autism is associated with impairments in attention, information processing, and memory [142]. Therapeutic interventions to support attention impairments for autism heavily rely on the use of physical objects and methods using stimulus-response-reward techniques to teach basic skills such as attention, compliance, and imitation [143]. However, most of the children with autism find task repetition boring and frustrating, and they do not find real objects appealing, as a consequence they might spend a lot of time "off-task" and may be unable to sustain their attention during therapy. Strategies caregivers use to maintain the attention of children include the annotation of text on top of physical objects [26] and the use of verbal and physical prompts and rewards. During each trial, teachers struggle to keep children with autism on task (e.g., engaged in the therapy and looking at the teacher), so teachers have to invest a lot of effort in helping children with autism during discrimination training.

To understand the characteristics of the visual support teachers use during discrimination therapies and to supplement our understanding from our literature review, we conducted a qualitative study at Pasitos—a specialized clinic where around 23 psychologists-teachers attend to close to 60 low-functioning children with autism. We conducted a total of 44 interviews, 111 h of passive observation, and several participatory design sessions to design, MOBIS (Fig. 6.2, Ref. [72]).

MOBIS is an AR system that enables teachers to superimpose digital content, including text, audio-recorded messages, and visual shapes (e.g., circles) on top of physical objects used during cognitive training. First, teachers use their android tablet to create a database of images by uploading photos of the objects used during therapies and associating digital content

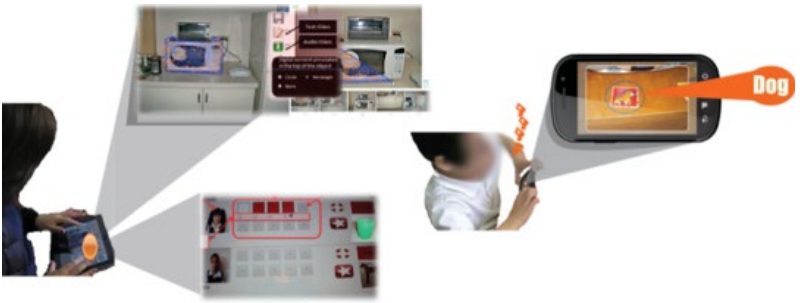


FIGURE 6.5 Top left: Teachers and children with autism from Pasitos using MOBIS (top left). Bottom left: A teacher, uploading photographs, tagging objects, and monitoring children with autism's responses. Right: A student using the smartphone to discover digital content on top of a paper-based card of a dog.

in the form of "tags" that will be later discovered by children (Fig. 6.5, top left). Then, teachers select from the tablet the object children need to discriminate and monitor their responses on each trial (Fig. 6.2 bottom left). Children later lay over their smartphone on top of a physical object using it as a "visor" to uncover the digital content tagged on top of the object (Fig. 6.5 right). To recognize objects, we used the SURF algorithm to extract features from images.

We deployed MOBIS in three different classrooms at Pasitos with 12 low-functioning children with autism (e.g., between the ages of 3 and 8 years $m = 5.08$, SD = 0.9) and 7 teachers (Fig. 6.6). Our study had two conditions: during *predeployment* (2 weeks) participants conducted their therapies using traditional paper-base visual supports; whereas during deployment (5 weeks), participants conducted their therapies using MOBIS. Researchers video recorded children with autism and teachers during therapies and conducted weekly interviews with the teachers across each study phase. The total time of observation was just under 54 h, and interviews lasted about 30 min ($m = 0:43:10$, SD = 01:10:05). We asked teachers about their experiences when using MOBIS and themes related to adoption. We also asked teachers about the potential areas impacted by MOBIS including engagement, attention, and behavior. Data analysis followed a mixed-method approach.

We found that MOBIS increased the time children with autism remain "on task" (baseline: 00:17:15; deployment: 03 :12:47; p[27] = 0.003). Children with autism were more motivated to conduct the therapy when using MOBIS than in baseline:

[27] The p-value or calculated probability is the estimated probability that is used for testing a statistical hypothesis. For our case, we defined a significance level of 5% meaning that a p-value ≤ 0.05 indicates that the results are statistically significant.

FIGURE 6.6 Participants across study phases. A student with autism attending to an object discrimination lesson using traditional paper-based visual supports (left), Pasitos staff during a participatory design session (center), and children with autism using MOBIS during an object discrimination lesson (right).

> Children with autism now enjoy the therapy [referring to the use of MOBIS]. They used to be apprehensive [when using the paper-based visual supports] but now they are more proactive and engaged in the therapy.

Children with autism improved their *selective attention*[28] (before—MOBIS: 00:01:05; using—MOBIS: 00:06:18; and after—MOBIS: 00:00:58, $p=0.0002$), as they were engaged in the therapy, even when something could potentially distract them (e.g., yells and noises in the environment).

> Children with autism were not distracted trying to get something near them, or seeing what happens to a classmate, or attending to some random noise, children with autism just get [MOBIS] and start to focus on the therapy (Adriana, teacher).
>
> For instance with [a child with autism with severe behavior problems], I struggle a lot …, he regularly makes big temper tantrums, […] but when using MOBIS I saw him very interested, he wanted to keep working, he didn't want to move from the therapy area, this was awesome! (Analí, teacher).

We also found MOBIS improved children with autism's *sustained attention*[29] increasing the time they remain consecutively "on task" (before—MOBIS: 00:00:41; using—MOBIS: 00:04:31; and after—MOBIS: 00:00:51, $p=0.0005$, Fig. 6.6).

> Before [MOBIS] some children with autism would not even listen to us even though we repeatedly called out their names. Now [with MOBIS] because of the smartphone, the system, the sounds, the visual stimulus, they are more engaged in the therapy and remain more time 'on task' (Bella, teacher).

Overall, MOBIS caught children with autism's attention in a simple and effective way, with the ambiguous role of presenting plenty but unobtrusive

[28] Selective attention refers to the ability of individuals to stay "on task" even when a distraction is present.

[29] Sustained attention is the amount of time individuals can concentrate on a task without becoming distracted—in other words, the ability to sustain selective attention.

stimuli due to the capabilities of AR technology that prove to be effective for attention management, and particularly for this population. These results highlight the importance of mimicking current practices for attention management as features when designing AR technologies.

Altogether, this case study illustrates AR can be effective in supporting therapies for children with autism. The augmented information superimposed over the objects children use during therapies induces positive changes in their attention.

All in all, this body of work shows tangible computing and AR tools are appropriate technologies to support cognitive training and facilitate the use of visual supports. These technologies could be integrated with other pervasive technologies appropriate for the use inside the classroom, especially with capture and access tools, to help identify when the student with autism is either "on task" or "off task" during therapies. Such an environment may also help teachers to automate the capturing of contextual information relevant to attention to evaluate children's progress. In this regard, it would be useful to design algorithms for automatic recognition of attention and engagement.

6.4 SERIOUS GAMES AND EXERGAMES TO SUPPORT MOTOR IMPAIRMENTS

Exergames combine games and exercise together [144, 145], and the player has to mimic the game activity on the screen through avatars [146]. The use of both avatars and movement increases exercise motivation [37]. Exergames are especially useful when running on interactive surfaces to support children with autism's social and physical interaction [147–149]. Interactive surfaces are of particular importance to support motor impairments, because they provide a natural interaction enabling individuals to appropriately use gross motor movements when playing. In addition, the ability of exergames to sustain engagement could positively impact therapy adherence.

6.4.1 Supporting the Practicing of Motor Coordination Exercises

Exergames are appropriate to help children practicing motor skills [37]. In the literature, several projects have investigated how exergames support the development of motor skills (e.g., arrow attack [150]) for different populations (e.g., children with cerebral palsy[30] [84], older adults [151], stroke patients [152], and autism [86–88]). Exergames have been used in rehabilitation therapies due to their capability to engage patients during motor therapies—especially individuals with Parkinson's—as motor skills are of vital importance for this population [85]. For example, several projects in the pervasive healthcare literature have proposed different therapeutic exergames to support balanced therapies [153] or the practice of upper limb movements ([151], see Case Study 6.3 [105]).

[30] Cerebral palsy is brain injury and a group of permanent movement disorders including poor coordination, stiff and weak muscles, trouble swallowing or speaking, and tremors.

Research in exergames for children with motor problems is mainly focused in support of motor development in children with cerebral palsy [84, 154, 155]. For example, Hernandez *et al.* [156] designed and developed an exergame that enables children with cerebral palsy to use a rehabilitation bicycle to control an avatar riding a unicycle while carrying a tray full of eggs in each hand. Other studies explore how to design and evaluate exergames for supporting the motor problems faced by children with autism or other related disorders. For example, Astrojumper [157] is an exergame motivating children with autism to exercise and practice motor skills. These projects demonstrate exergames that are appropriate interfaces to support individuals with cognitive disabilities when practicing upper limb movements.

Other projects have investigated how to support the practicing of lower limb movements (e.g., iGameFloor [158], Fish game [159], and Stomp [149]). For example, Stomp [149, 160] is an interactive floor including a set of educational open-ended games to support individuals with cognitive disabilities when learning maths and promote the appropriate gross motor skills development. A 2-weeks deployment study of the use of Stomps shows children with cognitive disabilities using exergames on interactive floors benefit from exercise, socialization, and cognition [149]. Similarly, Hunting Relics [161] is a collaborative interactive floor exergame that enables high-functioning children with autism collectively exercise like in a typical physical circuit training (Fig. 6.7). In Hunting Relics, children could practice different motor exercises to help two scouts when hunting for their father's lost relics. To get the relics, the scouts should avoid obstacles practicing basic eye–foot coordination exercises, like jumping, stomping, or walking in a straight line. The results of a deployment study show Hunting Relics enables the same physical activation as a traditional exercise circuit, augments existing exercise routines, maintains children engaged in long term, and persuades young children to discover new collaboration strategies to support exercising.

Still, major sociotechnical research questions must be addressed to make therapeutic exergames usable and effective. From a technical standpoint of view, we must first understand what body movements can be detected through automated machine-based methods and are appropriate to be represented through animation and how best to capture, model, and translate these movements. Open questions remain as to what are the best recognition algorithms to better monitor coordination exercises in real time and provide an appropriate feedback to individuals with cognitive disabilities when playing with exergames. From a persuasive standpoint of view, open questions remain to investigate if exergames are better solutions to encourage behavior change and their feasibility and efficacy of delivering therapeutic instruction via technological means. This will require a long-term study using persuasion theories to design better behavior change interventions.

6.4.2 Sensory Motor Integration

Researchers have also explicitly addressed how exergames might engage individuals with cognitive disabilities for therapeutic advantage during sensory motor integration therapies. Many individuals with autism, especially individuals with autism and attention-deficit hyperactivity disorder (ADHD), have sensory processing disorders [30].

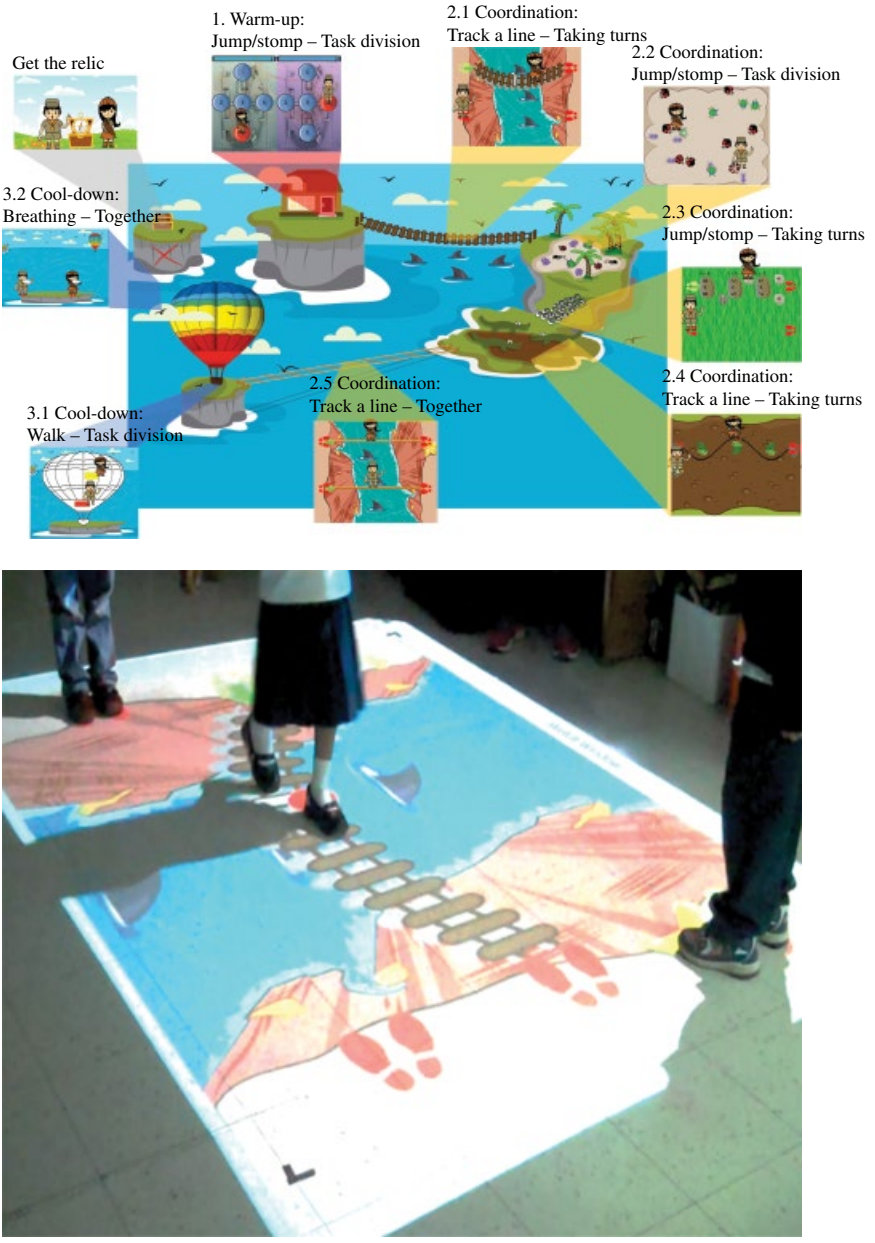


FIGURE 6.7 Hunting Relics: (top) Screenshots of the coordination challenges available in Hunting Relics and (bottom) children using the interactive floor and Hunting Relics.

People with sensory processing disorders lack appropriate sensory integration capabilities (i.e., "*the ability to use the neurological process to organize sensation from one's own body and the environment, thus making it possible to use the body effectively within the environment*" [162]). Individuals with sensory processing disorders are often hyper- or hyposensitive to light, sound, and/or touch [163] and experience poor body and movement awareness [164] that lead to atypical body interactions.

To support individuals with sensory processing disorders, the pervasive healthcare literature has researches on how exergames could augment music therapy [165] and motor-sensory integration therapeutic interventions (e.g., MEDIATE [148] and Sensory Paint [166, 167]). Previous work has demonstrated how tangible user interfaces [71, 140, 168] support the visual and auditory system of individuals with sensory processing disorders, particularly children with autism. Specialized devices that enhance musical therapies by providing additional visual [168, 169] and auditory [71, 170, 171] stimuli have been shown to reduce the sensory processing symptoms of sensory processing disorders. For example, Reactable [168, 169] is a tabletop surface used as a musical instrument. Users place wooden cubes emitting different sound effects, and when connected Reactable plays different melodies or sound patterns. In a deployment study involving children with autism, Reactable improved the social skills of participants [137]. Other systems have also been designed to combine sounds with visual stimuli [140, 162, 172, 173]. The resultant sensory stimulation of these systems provides platforms for children's self-reflection [173], engagement in the activity [174], self-directed activity [140], and language usage [162]. Other techniques integrate physical and digital media, including commercial devices (e.g., livescribe[31] running Tap and Play [140]) and the development of self-crafted devices to provide various means of interaction (e.g., OnObject [162], T3 [71], and Tobobo [171]).

Kinect games have been shown to be therapeutically beneficial for individuals with cognitive disabilities and hold their attention, reducing the need for therapist intervention [175]. At the same time, the Kinect has been shown to be useful in detecting a variety of user actions particular to individuals feeling restless or with fidgeting problems, such as hand-flapping, and other stereotypical [176] and expressive movements [177]. Research exploring new models for body-based interactions has explored how to support individuals to gain body awareness [172].

Projects exploring body-based interactions primarily focus on helping nonverbal individuals use their body to express themselves by redirecting users' attention toward their own bodies. For example, MEDIATE [148] is a body-based interactive environment designed to stimulate the creativity of nonverbal children with autism through vibrotactile, visual, and auditory stimuli generated in real time. MEDIATE includes sensors to obtain information from the child's activities, a vision system, and projection screens. When interacting with MEDIATE, children control the digital information displayed on the surface (e.g., leaves and snow) by moving their bodies. During a laboratory study of the use of MEDIATE, children enjoyed using movement for self-expression. While MEDIATE allows children to express themselves through

[31] http://www.livescribe.com.

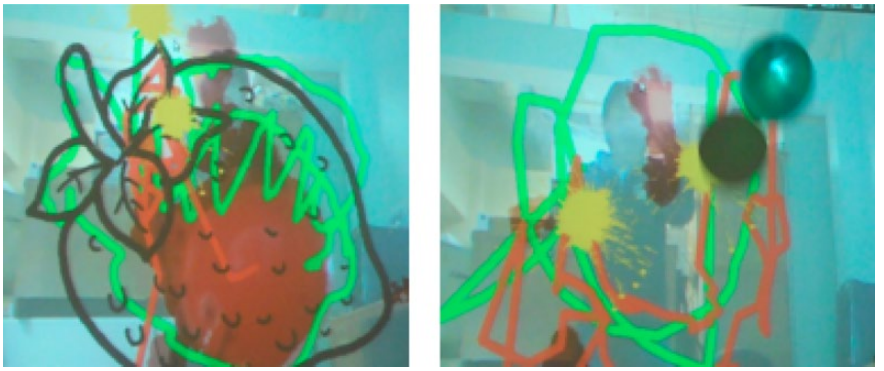


FIGURE 6.8 A user playing with the coloring book (left) and decorating paintings with "splashes" (right).

their body movements, it does not fully support proprioception[32] [138]. Similarly, BodyPaint [178] is an interactive system through which users paint with their bodies. However, BodyPaint assumes users have enough awareness of their own bodies to use them for painting. This might be inappropriate for individuals with poor body awareness who cannot control their own body, much less use it to manipulate the digital information showing on a large display.

Building upon this body of work, we have explored how to combine movement-based and tangible interactions to increase body awareness and support sensory integration therapies for children with autism. Sensory Paint [166, 167] is an exergame using an interactive painting tool, showing a superimposed reflection of the user projected onto a canvas or wall to highlight the actions of the user. The color of the user's reflection changes from red to green to demonstrate the proximity of the user to the surface. Rubber balls held by the user are detected by the system and act as paintbrushes of various sizes, textures, and colors, allowing for the painting of lines (Fig. 6.8, left). Users can either draw in a free form mode or use a template. Users can also splash color on the surface by throwing the rubber ball (Fig. 6.8, right). Finally, to complete the multisensory experience, sounds are played in connection with ball movement. The results of two empirical studies of the use of SensoryPaint show that this technology supports sensory integration, including stimulus sensitivity, body awareness, motor functioning, attention, and engagement.

In addition to the challenges associated with the multimodal features exergames to support sensory motor integration, most activities conducted in such an environment are open-ended. As a consequence, sensory therapists frequently borrow strategies from occupational therapies to provide individuals with cognitive disabilities with goal-oriented tasks to maintain their attention [164]. In contrast, with open-ended tasks that offer a more subtle stimulation, occupational therapies can increase engagement with sensory therapies by giving children a sense of control over the

[32] Proprioception is one's own sense of the relative position of neighboring parts of the body and strength of effort being employed in movement.

stimulation and a sense of purpose when interacting with stimuli [179]. These show the importance of combining open-ended and task-oriented interaction modalities to give users the opportunity to more freely personalize exergames to their needs and uncover other potential practices mediated by the technology, while sustaining engagement.

In the following text, we illustrate how an exergame we developed is successful in supporting the motor skills therapeutic interventions of children with autism.

Case Study 6.3 An Exergame to Support Children with Autism Practicing of Motor Skills

Children with autism commonly exhibit impairments in visual motor integration particularly gross motor eye–body coordination and fine motor eye–hand coordination. Designing exergames to meet the needs of children with autism is not an easy task. The present case study showcases the design space of exergames supporting children with autism when practicing eye–body coordination exercises during motor therapeutic interventions.[33]

We followed an iterative user-centered design methodology conducting a 6-month qualitative study, a 2-week qualitative study, and 13 participatory design sessions to uncover gameful mechanisms and motor coordination exercises that are appropriate to support therapeutic interventions for motor development. As a result of our iterative process, we designed and developed FroggyBobby [86–88], an exergame to support children with motor problems when practicing upper limb eye–body motor coordination exercises.

FroggyBobby uses a frog avatar whose tongue is controlled by children's arms. When using FroggyBobby, children have to practice with each specific arm the gross motor exercises our design team selected as important to help children develop age-appropriate motor skills (Table 6.3). The *basic gameplay* of FroggyBobby demands children move their left or right arm from up to down in a lateral way. Children start by tapping on the start button, located in the upper area of the screen (Fig. 6.9 left), and "swiping" their arm toward the end button located in the lower area of the screen to catch as many flies in the path between the start and the end buttons (Fig. 6.9 right). For each level, children alternate between using the right and left arm.

To evaluate the impact of FroggyBobby, we conducted a 3-week deployment study at Pasitos.[34] Researchers installed FroggyBobby in Pasitos and gave a training lesson to the children with autism and their teachers participating in the study. A week was used to calibrate the exergame. After that,

[33] Therapeutic interventions for motor development of children with motor problems include physical therapy (e.g., for practicing motor exercises to improve gross motor coordination), occupational therapy (e.g., for the development of activities that requires gross and fine motor coordination), among others [180, 181].

[34] A specialized school clinic, where 15 psychologist-teachers attend around 50 low- and mid-functioning children with autism, http://www.pasitos.org/.

TABLE 6.3 Set of Gross Motor Exercises Proposed During Participatory Design Sessions and Included in FroggyBobby as Game Controllers and Motor Exercises Adequate for Motor Skills Development

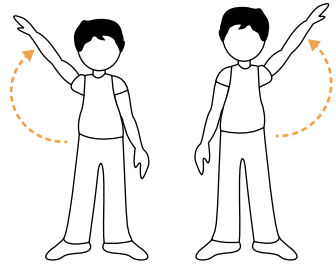
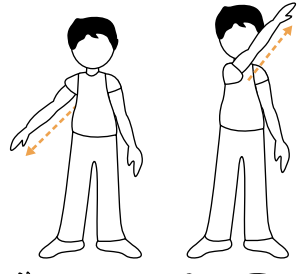
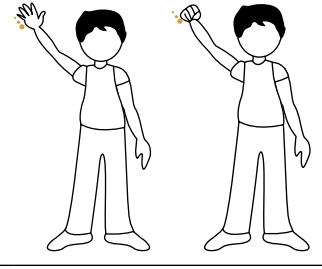
Exercises	Description	Example
Strength and motion control exercises	Children move each arm from up to down "swiping" in a lateral way drawing a half circle. This group of exercises can improve "arms strength/flexibility and arc of movement" and impact in the execution of different ADL including reaching an object, hair washing, bathing, and drying with a towel.	
Cross-lateral exercises	Children move each arm from up to down crossing from left to right. This group of exercises can improve "brain connection and body coordination" and impact in different motor activities that require performing cross-lateral movements (e.g., picking an object, washing the body, and tying shoelaces)	
Coordination and visual spatiality exercises	Children maintain each arm in a steady position and perform a fine gross movement (e.g., perform a grip movement to simulate reaching an object). This group of exercises can improve "arms coordination, spatial orientation, and spatial dimensions" and impact in the accuracy of goal-directed movements, dressing activity, or in the manipulation of different objects.	

FIGURE 6.9 First level of FroggyBobby: (left) A child moves the right arm to reach the start button (top right button) and Bobby extends its tongue to reach the start button and (right) a child moves the right arm to reach the end button (lower right button) to catch as many flies between the start and the end button. Bobby brings its tongue back to its mouth and swallows all caught flies.

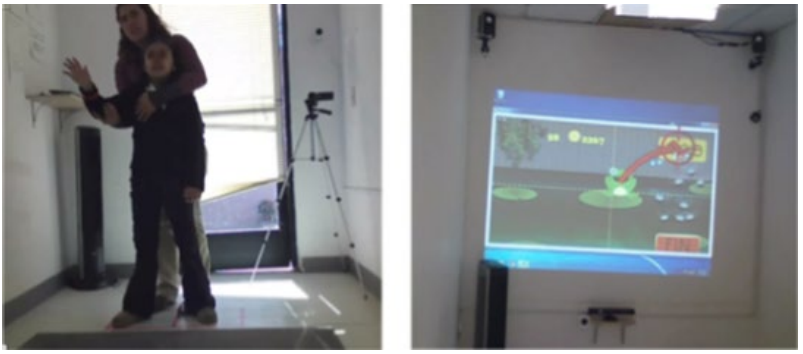


FIGURE 6.10 A child playing with FroggyBobby. (left) A screenshot of the front camera showing a child playing with FroggyBobby while the teacher prompts her physically when practicing different motor exercises. (right) A screenshot of the back camera overlooking the multimedia projection of FroggyBobby exergame being played by a child.

children with autism played FroggyBobby for the following 2 weeks. During each game session, FroggyBobby was played for about 07:53 min on average per child (SD = 00:04:32).

Overall, children and teachers perceived FroggyBobby (Fig. 6.10) "easy to use" and "useful." All teachers commented they were enthusiastic with the use of FroggyBobby. They perceived the exergame as a useful tool, particularly supporting low-functioning children with autism when practicing motor skills while promoting positive emotions of children with autism.

> [With FroggyBobby, children with autism] are practicing their movements while they are playing, they are learning to do something that is difficult for them… (A2, teacher).

Teachers reported that over the course of using FroggyBobby, children with autism required less prompts and less instructions related to how to play FroggyBobby. Although teachers continuously helped children when using their arms to reach the buttons, teachers explained that they perceived the number of prompts decreased over the four sessions of use.

> At the beginning, we had to help children with autism to raise [their arms], but then, my hand felt like a magnet, like I did not have to [use a lot of strength] to help children with autism raise their arms, it was like their hand was following mine. They were doing the motor exercise by themselves. Now I only prompt them by saying: 'raise your arm higher' (A1, teacher).

Overall, FroggyBobby was well accepted, and teachers perceived and foresaw further impact in motor functioning, socialization, and cognition. FroggyBobby served as a true support tool leveraging current practices around motor therapeutic interventions and supporting the practicing of motor coordination exercises while sustaining the attention of children with autism.

Altogether, this body of work showcases different reflections about exergames from design, technical, and clinical standpoints. From a design standpoint, we learned the importance of using appropriate stimuli in the design of exergames for individuals with cognitive disabilities. Our results indicate that it is important to include customizable prompts and rewards in the exergame. Although there are several projects exploring how to support the practicing of gross motor skills, little has been said if pervasive technology is appropriate to support fine motor development. This leaves new challenges about designing appropriate fine motor coordination exercises that could be supported with pervasive technology and more particularly with exergames. From a technical standpoint of view, more research is needed in order to investigate how reliable current tracking solutions are to support fine motor coordination exercises. These technologies should be reliable enough to the frustration of users when practicing movements that are not properly detected. Finally, from the therapeutic standpoint, it would be interesting to investigate the long-term impact and skills generalization due to the use of exergaming technology.

6.4.3 Assisted Navigation and Wayfinding Support

Mobility can be a substantial challenge for people with cognitive disabilities attempting to live independently [182]. These challenges can include problems like working, driving, or taking public transportation. Beyond outdoor mobility, others face orientation issues when navigating indoors, even in well-known spaces such as a home. To support wayfinding, many research projects have worked toward helping disoriented people with sensory impairments to navigate—in both indoor and outdoor environments—using a variety of technological solutions.

The Intelligent Mobility Platform is a walker-based device that uses a laser beam range finder, a mobile device, and a navigation software to orient a person in the proper direction [183]. In the Indoor wayfinding project, researchers designed an

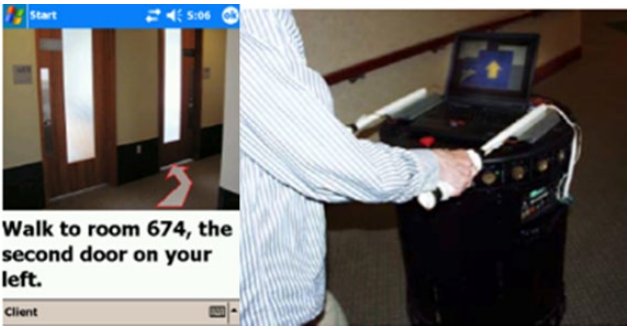


FIGURE 6.11 The Indoor Wayfinding System showing directions (left); and the Robotic Walker (right).

interface for mobile devices to send directions and prompts to the user. Images, audios, and text messages were combined to provide prompting and directions. For example, image-based directions include photos of the outlined area overlaid with direction arrows. Others have explored the use of tour-guide robots as a natural way for providing mobility guidance [183, 184], facilitate walking (e.g., the walk training assist robot) adjusting to individual's weight, and support the rehabilitation training (e.g., the balance training assistant).[35] Nowadays, robots have proven to be very beneficial in helping individuals with restricted mobility to live with a greater independence.[36] The Robotic Walker [183] includes a robot that physically guides the elderly within an assisted living home. In this project, a Nomad mobile robot was equipped with an omnidirectional drive and software with capabilities for planning paths, tracking people, and avoiding collision. The robotic walker has a touch interface for receiving commands from the individual. Simple directions in the form of an arrow are shown in a mounted display. This system provides not only physical support for walking but also wayfinding support. In all of these projects, a variety of pervasive computing technologies are used together in a sort of assistive assemblage to provide the level of support required. Thus, pervasive health applications not only provide motivation for technologists to develop these new systems but also a means for understanding the potential for and evaluating the impact of these vast and varied collection of technologies (Fig. 6.11).

Indoor wayfinding is a challenge for many populations beyond the elderly, including the visually impaired [185] and children with autism. However, thus far, they have largely been tested with groups of people who might already be using other assistive devices. This approach avoids the stigmatization and other social issues that might come with using such systems. For example, if an elderly person is already using a walker, adding some computation to this physical platform does not substantially change the overall look of it. However, a younger able-bodied person with a

[35] Toyota healthcare robots, http://mashable.com/2011/11/01/toyota-healthcare-robots/.
[36] http://www.networkworld.com/news/2009/041609-robots-will-aid-in-health.html.

brain injury might not wish to use a walker that would add further stigma. The design of less-obtrusive systems for those who may have different types of disabilities and therefore not be willing to wear or use such obvious assistive systems remains an open design challenge for the future.

In contrast to these indoor navigation systems [186], others have explored the use of auditory aids [187] and photograph-based navigation interfaces [188] for outdoor navigation guidance. For example, Opportunity Knocks is a mobile phone-based device using GPS and Bluetooth that learns the user's standard routes in the community to route an individual from their current location to a chosen destination [189]. Opportunity Knocks requires very little user input. Instead it relies on observed user history as a basis for predicting likely destinations and identifying novel and erroneous behavior. It alerts the person of a navigational error by making a knocking sound and subsequently recalculating the proper route.

In contrast, most other outdoor wayfinding systems have concentrated on assisting people while driving [186, 190–192]. Commercial applications for using GPS and mobile phone towers to triangulate location and provide directions are numerous and out of the scope of this chapter, which focuses on independent living for individuals with chronic health conditions. Wayfinding technologies can assist people with a variety of health impairments who may be disoriented but still need mobility both at home and in the outdoor environment. Research in this area has demonstrated that current technologies can enable limited assistance. However, more work is needed to improve accuracy when estimating location, particularly indoors and with the kind of precision required to support the activities of the chronically ill at home.

6.5 CONCLUSIONS

In this concluding section, we summarize major approaches that projects on pervasive health care have proposed to support individuals with cognitive disabilities. We then conclude with a discussion of future work, open challenges, and future trends.

6.5.1 Application Themes

For several years, numerous pervasive healthcare projects have been devoted in seeking the Holy Grail of pervasive health care: the "killer app" that will shift the healthcare model, as we know it. Indeed, one of the biggest challenges in pervasive health care involves applications and experiences—descriptions of the design or empirical study of applications that leverage Ubicomp devices and systems targeting some challenge associated to health.

The constant evolution of the wearable and mobile sensing platforms opens new possibilities for data acquisition that combined with wireless networks, consuming electronic technologies, and with capabilities for inferring behavioral data, which have the potential to change our lifestyle. The integration of data mining techniques and appropriate visualizations to create adequate displays will be paramount.

These innovative displays could help people collect personal relevant information to encourage self-reflection and potentially promote behavior change.

Exergames and tangible computing solutions should examine ways to personalize and customize natural user interfaces for therapeutic interventions across a wide variety of contexts and users. Natural interfaces include a wide variety of interaction styles between humans and computers but generally include those interaction modalities that are the most *natural* in terms of being familiar to the average person. These interfaces are intended to blend seamlessly into the everyday behaviors of individuals. Typical natural interfaces include the use of speech, pens, and gestures. The benefits to a "natural" approach to interfacing with computational devices are enormous. They have the potential for significant hands-free interaction, like Google glasses.[37]

Beyond the typical natural interfaces that use speech, pens, gestures, etc., paradigms are emerging that could push the limits of our imagination. For example, brain–computer interfaces (BCIs) could give users feedback about their own brain activity and adapt the smart environment accordingly—some exergames might turn off some features when individuals with cognitive disabilities start losing attention. However, open questions remain as to how BCIs could be integrated into existing pervasive healthcare prototypes to give users new controllers to manipulate digital and physical objects.

Recent advances in computer vision and audio recognition also make possible the creation of novel interfaces with innovative means for interaction. For example, blendable or flexible surfaces could let users more freely gravitate and manipulate digital objects displayed on the surface. The affordances of flexible surfaces invite users to grasp, push, or bend the surface, which in turn lets them discover multisensory interaction experiences.

6.5.2 Trends and Open Challenges

The trends in wearable and mobile sensing platforms on pervasive health care are increasing with recent advances and new proposal of body sensors, the incorporation of more sensors in smartphones, and the low cost and accessibility of heterogeneous devices to potential users.

Although the number of research using wearable and mobile sensing platforms for pervasive health care is increasing, there are still several challenges to be addressed. From a clinical point of view, there remains many questions as to how patient data acquired through mobile sensing applications could be used to help clinicians provide an early diagnosis and uncover behavior patterns and causes of intellectual disabilities.

Second, from a technical point of view, there are other open challenges around the proposal of novel algorithms to improve the detection and prediction of context and manage uncertainty. For many pervasive health applications to perform at their

[37] https://www.google.com/glass/start/.

highest capabilities—or in some cases to work at all—some amount of context must be sensed, recognized, and presented to the system or its users. However, sensors and recognition algorithms are rarely error-free. Understanding how to present and to manage uncertainty remains an open area of research. The presence of uncertainty information in pervasive healthcare applications represents one of the main challenges and one of the greatest barriers to the long-term success of pervasive healthcare applications. Open questions to increase adoption of these technologies and their integration into pervasive computing systems involve solving challenges around invisibility, sensitivity of sensors, battery life, data collection, safe data transfer, compatibility, and bandwidth.

For exergames, pervasive technology should enable a simple and natural interaction model. A lot of natural user interfaces (NUIs) and interactive surfaces for exergames use vision-based algorithms demanding complex setup for the environment to appropriately calibrate the system to the conditions of the environment, and they need to be appropriately calibrated to provide an adequate user experience. The exergames need to include a reliable tracking mechanism for detecting different movements and gestural interactions. Having a wrong interaction model could generate frustration when fine movements are not properly detected. Also, an important tendency is to include collaboration mechanisms to enable the collective use of exergames. Collective use supports engagement, promotes imitation, and enables more users to benefit from the use of the system on a given day. The main challenges around collaboration include balancing game challenges and player skills, because some users will have more capabilities than others. Mechanisms like handicapping from golfing and personalization are currently scarce in the literature and much needed.

Similarly, one of the most important decisions when developing tangible computing AR services is to enable a real enhancing experience while keeping the interaction model as simple and natural as possible. For example, the use of MOBIS leaves open questions for creating a more suitable "AR visor" to help individuals with cognitive disabilities with the manipulation and discovering of AR services. Overall, when selecting the most simple and accurate tool to provide AR services, it is important to consider the characteristics of the user and the efforts to setup the device to be used in real-life situations.

These projects have highlighted how far we have come in providing support for a variety of activities that seemed impossible just a few decades ago. However, we are just at the beginning of exploring this space. Although, current deployment studies have shown multiple health-related benefits to their potential users, there is not much evidence of their benefits in long term. Long-term deployment studies exploring questions around persuasion and studying healthcare efficacy over long term are currently needed. These will open up questions around the design of behavior change interventions mediated by pervasive healthcare technology, exploring issues emerging after the technology novelty effect wears off.

The use of this technology is moving healthcare services outside hospitals and clinics into the patient's home. This in turn could potentially decrease hospital costs, but the deployment of some of these technologies at home is still expensive. Some

pervasive healthcare prototypes demand to purchase expensive equipment or modify home premises to make possible the deployment of such technology. The viability of projects based on a vast number of sensors combined with other techniques will depend significantly on the cost of the devices, replacement period, and automated network configuration. Not only will some deployed pervasive healthcare prototypes eventually become legacy systems once the host hardware's novelty expires, but also the increasing installation of new software and hardware will add an extra maintenance and integration burden on users. Scalability and costs are challenges that hamper the deployment and adoption of pervasive healthcare technologies in real-world settings. In particular, new standards for communication and design across a wide array of medical devices and legacy health information systems might help to reduce challenges and associated costs when deploying new pervasive and Ubicomp technologies in real settings.

6.5.3 Methods and Tools

There is a gap between the methods used in clinical research and computer science. For our particular case, the planning of our studies involved constant sense making in collaboration with clinicians, teachers, and psychologists working at healthcare institutions to make sure there is an agreement of expectations between clinicians and computer scientists. As a result, we slightly adapted our data collection techniques to provide additional value to our clinical partners and ensure more active engagement. These show study planning must be participatory as both clinicians and researchers participating in the study value empirical measurement.

Our experiences show that the deployment of pervasive healthcare solutions becomes an opportunity to gather verifiable and quantifiable data in the form of a heterogeneous massive database containing videos, audios from interviews, photos, and data from sensors. However, capturing too much data could make data analysis unbearable. This calls for new tools and techniques to appropriately label and segment the data being captured and to more easily help researchers isolate those data segments that are relevant to the understanding of the phenomenon being studied and significantly reduce the burden and workload associated with data capturing and analysis. From a technical standpoint, this heterogeneous database could be useful to train and test classifiers and models for predicting behavior.

Finally, it's not clear when it's appropriate to start redesigning already deployed prototypes, or when it's suitable to deploy new prototypes or a new version of existing ones. It takes time for the novelty effect to wear off and for participants to start feeling comfortable using the prototypes; this heavily limits the trend of building "semiworking" technologies and deploying them as quickly as possible because frequent technology updates could be very disruptive to existing practices related to adoption. So, we must find new methods to promptly and appropriately integrate incremental, transformative innovations into existing smart environments [193].

ACKNOWLEDGMENTS

We thank the participants of the studies and the Pasitos school-clinic for their support. This work was funded through Dr. Monica Tentori's Microsoft Fellowship gift, UC Mexus and LACCIR grant. The following faculty, postdocs, and children with autism (listed in alphabetic order), helped develop the technologies presented in the three case studies described in this chapter: Karina Caro, Luis Castro, Jesus Favela, Daniel Garcia-Rosas, Gillian R. Hayes, Moises Perez, Eduardo Quintana, Alejandro Rangel, Kate Ringland, Ivan Zavala, and Rodrigo Zalapa.

REFERENCES

[1] American Psychiatric Association., Diagnostic and Statistical Manual of Mental Disorders, (DSM-5®). 2013: American Psychiatric Publishing, Washington, DC.

[2] Mackelprang, R.W. and R.O. Salsgiver, Disability: A Diversity Model Approach in Human Service Practice. 2nd Edition, 2009: Lyceum Books, Chicago. p. 467.

[3] Levy, R., Aging-associated cognitive decline. International Psychogeriatrics, 1994. 6: 63–68.

[4] Brisenden, S., Independent living and the medical model of disability. Disability and Society, 1986. 1(2): 173–178.

[5] Deegan, P.E., The independent living movement and people with psychiatric disabilities: taking back control over our own lives. Psychiatric Rehabilitation Journal, 1992. 15(3): 3–19.

[6] Giangreco, M. and S. Broer, Questionable utilization of paraprofessionals in inclusive schools: are we addressing symptoms or causes? Focus on Autism and Other Developmental Disabilities, 2005. 20: 10–26.

[7] Galasko, D., D. Bennett, M. Sano, C. Ernesto, R. Thomas, M. Grundman, and S. Ferris, An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. Alzheimer Disease & Associated Disorders, 1997. 11: 33–39.

[8] Greiner, P.A., D.A. Snowdon, and F.A. Schmitt, The loss of independence in activities of daily living: the role of low normal cognitive function in elderly nuns. American Journal of Public Health, 1996. 86(1): 62–66.

[9] Giangreco, M., M.F. Giangreco, S. Edelman, T.E. Luiselli, and S.Z. MacFarland, Helping or hovering? Effects of instructional assistant proximity on students with disabilities. Exceptional Children, 1997. 64: 7–18.

[10] Hayes, G.R., L.M. Gardere, G.D. Abowd, and K.N. Truong. CareLog: A Selective Archiving Tool for Behavior Management in Schools, in Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems, Florence, Italy, April 5–10, 2008. ACM, New York. p. 685–694.

[11] Schulz, R. and L.M. Martire, Family caregiving of persons with dementia: prevalence, health effects, and support strategies. The American Journal of Geriatric Psychiatry, 2004. 12(3): 240–249.

[12] Wolff, J.L. and J.D. Kasper, Caregivers of frail elders: updating a national profile. The Gerontologist, 2006. 46(3): 344–356.

[13] Garwood, S., Cardiac surgery-associated acute renal injury: new paradigms and innovative therapies. Journal of Cardiothoracic and Vascular Anesthesia, 2010. 24(6): 990–1001.

[14] Ong, V.H. and C.P. Denton, Innovative therapies for systemic sclerosis. Current Opinion in Rheumatology, 2010. 22(3): 264–272.

[15] Alankus, G. and C. Kelleher. Reducing Compensatory Motions in Video Games for Stroke Rehabilitation, in Proceedings of the 30th ACM Annual Conference on Human Factors in Computing Systems, Austin, Texas, May 5–10, 2012. ACM, New York. p. 2049–2058.

[16] Alankus, G., R. Proffitt, C. Kelleher, and J. Engsberg, Stroke therapy through motion-based games: a case study. ACM Transactions on Accessible Computing, 2011. 4(1): 3.

[17] Gómez, J., G. Montoro, P.A. Haya, X. Alamán, S. Alves, and M. Martínez, Adaptive manuals as assistive technology to support and train people with acquired brain injury in their daily life activities. Personal and Ubiquitous Computing, 2013. 17(6): 1117–1126.

[18] Hartley, S., D. Sikora, and R. McCoy, Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. Journal of Intellectual Disability Research, 2008. 52(10): 819–829.

[19] Sadowsky, C.H. and J.E. Galvin, Guidelines for the management of cognitive and behavioral problems in dementia. The Journal of the American Board of Family Medicine, 2012. 25(3): 350–366.

[20] Plötz, T., N.Y. Hammerla, A. Rozga, A. Reavis, N. Call, and G.D. Abowd. Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, September 5–8, 2012. ACM, New York. p. 391–400.

[21] Rothbaum, B.O., E.A. Meadows, P. Resick, and D.W. Foy, Cognitive-Behavioral Therapy. 2000. Guilford: New York.

[22] Williams, G., L.A. Pérez-González, and A. Muller, Using a combined blocking procedure to teach color discrimination to a child with autism. Journal of Applied Behavior Analysis, 2005. 38(4): 555–558.

[23] Cohen, M.D. and D.L. Sloan, Visual Supports for People with Autism: A Guide for Parents and Professionals. 2008. Woodbine House, Bethesda.

[24] Hirano, S., M. Yeganyan, G. Marcu, D. Nguyen, and G.R. Hayes, vSked: Evaluation of a System to Support Classroom Activities for Children with Autism, in Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, April 10–15, 2010. ACM, New York. p. 1633–1642.

[25] Escobedo, L., D. Nguyen, L. Boyd, A. Rangel, D. García, S. Hirano, M. Tentori, and G. Hayes, MOSOCO: A Mobile Assistive Tool to Support Children with Autism Practicing Social Skills in Real-Life Situations, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), Austin, TX, May 5–10, 2012. ACM, New York. p. 2589–2598.

[26] Hayes, G.R., S. Hirano, G. Marcu, M. Monibi, D.H. Nguyen, and M. Yeganyan, Interactive visual supports for children with autism. Personal and Ubiquitous Computing, 2010. 14(7): 663–680.

[27] Fournier, K.A., C.J. Hass, S.K. Naik, N. Lodha, and J.H. Cauraugh, Motor coordination in autism spectrum disorders: a synthesis and meta-analysis. Journal of Autism and Developmental Disorders, 2010. 40(10): 1227–1240.

[28] Jasmin, E., M. Couture, P. McKinley, G. Reid, E. Fombonne, and E. Gisel, Sensori-motor and daily living skills of preschool children with autism spectrum disorders. Journal of Autism and Developmental Disorders, 2009. 39(2): 231–241.

[29] Shumway-Cook, A. and M.H. Woollacott, Motor Control: Theory and Practical Applications. Vol. 157. 1995: Williams & Wilkins, Baltimore.

[30] Ayres, A.J. and J. Robbins, Sensory Integration and the Child: Understanding Hidden Sensory Challenges. 2005: Western Psychological Services, Los Angeles, CA.

[31] Barsch, R.H., Achieving Perceptual-Motor Efficiency, A Space-Oriented Approach to Learning. Perceptual Motor Curriculum, Volume I. 1967: Special Child Publications, Seattle.

[32] Cratty, B.J., Physical Expressions of Intelligence. 1972: Prentice-Hall, Englewood Cliffs.

[33] Laszlo, J.I. and P.J. Bairstow, Perceptual-Motor Behaviour: Developmental Assessment and Therapy. 1985: Praeger, New York, NY.

[34] Revie, G. and D. Larkin, Task-specific intervention with children reduces movement problems. Adapted Physical Activity Quarterly, 1993. 10: 29–29.

[35] Schoemaker, M., A. Niemeijer, K. Reynders, and B. Smits-Engelsman, Effectiveness of neuromotor task training for children with developmental coordination disorder: a pilot study. Neural Plasticity, 2003. 10(1–2): 155–163.

[36] Valentini, N. and M. Rudisill, Motivational climate, motor-skill development, and perceived competence: two studies of developmentally delayed kindergarten children. Journal of Teaching in Physical Education, 2004. 23(3): 216–234.

[37] Yim, J. and T.C.N. Graham. Using Games to Increase Exercise Motivation, in Proceedings of the 2007 Conference on Future Play, Toronto, Canada, November 15–17, 2007. ACM, New York. p. 166–173.

[38] Goh, D.H., R.P. Ang, and H.C. Tan, Strategies for designing effective psychotherapeutic gaming interventions for children and adolescents. Computers in Human Behavior, 2008. 24(5): 2217–2235.

[39] Slocum, S.K., S.J. Miller, and J.H. Tiger, Using a blocked-trials procedure to teach identity matching to a child with autism. Journal of Applied Behavior Analysis, 2012. 45(3): 619–624.

[40] Weiser, M., The computer for the 21st century. SIGMOBILE Mobile Computing and Communications Review, 1999. 3(3): 3–11.

[41] Borriello, G., V. Standford, C. Narayanaswami, and W. Menning, Pervasive computing in healthcare. IEEE Computer Society, 2007. 6: 17–19.

[42] Saha, D. and A. Mukherjee, Pervasive computing: a paradigm for the 21st century. IEEE Computer, 2003. 36(3): 25–31.

[43] Weiser, M., The computer for the 21st century. Scientific American, 1991. 265(3): 94–104.

[44] Weiser, M., Some computer science problems in ubiquitous computing. Communications of the ACM, 1993. 36(7): 78–84.

[45] Weiser, M., The future of ubiquitous computing on campus. Communications of the ACM, 1998. 41(1): 41–42.

[46] Dey, A.K. and D. Estrin, Perspectives on pervasive health from some of the field's leading researchers. IEEE Pervasive Computing, 2011. 10(2): 4–7.

[47] Chiu, M.C., S.P. Chang, Y.C. Chang, H.H. Chu, C.C.H. Chen, F.H. Hsiao, and J.C. Ko. Playful Bottle: A Mobile Social Persuasion System to Motivate Healthy Water Intake, in

Proceedings of the 11th ACM International Conference on Ubiquitous Computing, Orlando, FL, September 30 to October 3, 2009. ACM, New York. p. 185–194.

[48] Gasca, E., J. Favela, and M. Tentori. Persuasive Virtual Communities to Promote a Healthy Lifestyle among Patients with Chronic Diseases, in Proceedings of the 14th international workshop, CRIWG 2008, Omaha, NE, September 14–18 2008. Springer, Berlin. p. 74–82.

[49] Grimes, A., R. Harper, and R.E. Grinter, Celbratory health technology. Journal of Diabetes Science and Technology, 2011. 5(2): 319–324.

[50] Grimes, A., D. Tan, and D. Morris. Toward Technologies that Support Family Reflections on Health, in Proceedings of the ACM 2009 International Conference on Supporting Group Work, Sanibel Island, FL, May 10–13, 2009. ACM, New York. p. 311–320.

[51] Hoysniemi, J., International survey on the dance dance revolution game. ACM Computers in Entertainment, 2006. 4(2): 1544–3574.

[52] Liu, L.S., S.H. Hirano, M. Tentori, K.G. Cheng, S. George, S.G. Park, and G.R. Hayes. Improving Communication and Social Support for Caregivers of High-Risk Infants Through Mobile Technologies. in Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, March 19–23, 2011. ACM, New York. p. 475–484.

[53] Bardram, J. and H.B. Christensen, Pervasive computing support for hospitals: an overview of the activity-based computing project. IEEE Pervasive Computer, 2007. 6(1): 44–51.

[54] Bardram, J. and T. Hansen. The AWARE Architecture: Supporting Context-Mediated Social Awareness in Mobile Cooperation, in Proceedings of the 2004 ACM conference on Computer Supported Cooperative Work, Chicago, IL, November 6–10, 2004. ACM, New York. p. 192–201.

[55] Bardram, J., T. Hansen, and Soegaard, M. AwareMedia: A Shared Interactive Display Supporting Social, Temporal, and Spatial Awareness in Surgery, in Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work, Banff, Alberta, CAN, November 4–8, 2006. ACM, New York. p. 109–118.

[56] Bardram, J., T.R. Hansen, M. Mogensen, and M. Soegaard. Experiences from Real-World Deployment of Context-Aware Technologies in a Hospital Environment, in Proceedings of the 8th International Conference on Ubiquitous Computing, Orange County, CA, September 17–21, 2006. Springer, Berlin/Heidelberg. p. 369–386.

[57] Bardram, J.E., H. Baldus, and J. Favela, Pervasive Computing in Hospitals. In Pervasive Healthcare: Research and Applications of Pervasive Computing in Healthcare, ed. I.P.H.R.a.A.o.P.C.i. Healthcare. 2006: CRC Press, Boca Raton. p. 49–78.

[58] Abowd, G.D. What Next, Ubicomp?: Celebrating an Intellectual Disappearing Act, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, September 5–8, 2012. ACM, New York. p. 31–40.

[59] Goodwin, M.S., M. Haghighi, Q. Tang, M. Akcakaya, D. Erdogmus, and S.S. Intille, Moving Towards a Real-Time System for Automatically Recognizing Stereotypical Motor Movements in Individuals on the Autism Spectrum Using Wireless Accelerometry, in Proceedings of the 2014 ACM International Joint Conference/Pervasive and Ubiquitous Computing (UbiComp '14), Seattle, WA, September 13–17, 2014. ACM, New York. p. 861–872.

[60] Hayes, G.R., D.J. Patterson, M. Singh, D. Gravem, J. Rich, and D. Cooper, Supporting the transition from hospital to home for premature infants using integrated mobile computing and sensor support. Personal and Ubiquitous Computing, 2011. 15(8): 871–885.

[61] Kientz, J., S. Boring, G. Abowd, and G. Hayes, Abaris: Evaluating Automated Capture Applied to Structured Autism Interventions, in Proceedings of the 7th International Conference on Ubiquitous Computing, Tokyo, Japan, September 11–14, 2005. Springer, Berlin, Heidelberg. p. 323–339.

[62] Lee, S.W. and K. Mase, Activity and location recognition using wearable sensors. IEEE Pervasive Computing, 2002. 1(3): 24–32.

[63] Marcu, G., A.K. Dey, and S. Kiesler, Parent-Driven Use of Wearable Cameras for Autism Support: A Field Study with Families, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, September 5–8, 2012. ACM, New York. p. 401–410.

[64] Oliveira, E., M. Benovoy, N. Ribeiro, and T. Chambe, Towards Emotional Interaction: Using Movies to Automatically Learn Users' Emotional States, in Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part I, Lisbon, Portugal, September 5–9, 2011. Springer-Verlag, New York. p. 152–161.

[65] Picard, R.W., Affective Computing. 2000: MIT Press, Cambridge.

[66] Picard, R.W., E. Vyzas, and J. Healey, Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. 23(10): 1175–1191.

[67] Goodwin, M.S., S.S. Intille, F. Albinali, and W.F. Velicer, Automated detection of stereotypical motor movements. Journal of Autism and Developmental Disorders, 2011. 41(6): 770–782.

[68] Lai, C.-F., S.-Y. Chang, H.-C. Chao, and Y.-M. Huang, Detection of cognitive injured body region using multiple triaxial accelerometers for elderly falling. IEEE Sensors Journal, 2011. 11(3): 763–770.

[69] Azuma, R., A survey of augmented reality. Presence: Teleoperators and Virtual Environments, 1997. 6(4): 355–385.

[70] Cheok, A.D., X. Yang, Z.Z. Ying, M. Billinghurst, and H. Kato, Touch-space: mixed reality game space based on ubiquitous, tangible, and social computing. Personal and Ubiquitous Computing, 2002. 6(5–6): 430–442.

[71] Escobedo, L., C. Ibarra, J. Hernandez, and M. Tentori, Smart objects to support the discrimination training of children with autism. Personal and Ubiquitous Computing, 2013. 17(8): 1485–1497.

[72] Escobedo, L., M. Tentori, E. Quintana, J. Favela, and D. García-Rosas, Integrating the physical and the digital world to increase the attention of children with autism. IEEE Pervasive Computing, 2014 (in press/forthcoming).

[73] Quintana, E., C. Ibarra, L. Escobedo, M. Tentori, and J. Favela. Object and gesture recognition to assist children with autism during the discrimination training, in Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3–6, 2012. Springer, Berlin. p. 14–36.

[74] Bimbrahw, J., J. Boger, and A. Mihailidis, Investigating the efficacy of a computerized prompting device to assist children with autism spectrum disorder with activities of daily living. Assistive Technology, 2012. 24(4): 286–298.

[75] Boger, J., J. Hoey, P. Poupart, C. Boutilier, G. Fernie, and A. Mihailidis, A planning system based on Markov decision processes to guide people with dementia through activities of daily living. IEEE Transactions on Information Technology in Biomedicine, 2006. 10(2): 323–333.

[76] Mihailidis, A., B. Carmichael, and J. Boger. The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. IEEE Transactions on Information Technology in Biomedicine. 2004. 8: 238–247.

[77] Beukelman, D. and P. Mirenda, Augmentative and Alternative Communication Supporting Children and Adults with Complex Communication Needs. 2005. Paul H. Brookes Publishing, Baltimore.

[78] Burke, R., D.R. Beukelman, and K. Hux, Accuracy, efficiency and preferences of survivors of traumatic brain injury when using three organization strategies to retrieve words. Brain Injury, 2004. 18(5): 497–507.

[79] De Leo, G. and G. Leroy, Smartphones to Facilitate Communication and Improve Social Skills of Children with Severe Autism Spectrum Disorder: Special Education Teachers as Proxies, in Proceedings of the 7th International Conference on Interaction Design and Children, Chicago, IL, June 11–13, 2008. ACM, New York. p. 45–48.

[80] Gillett, T., Kellen, M., Ball, L. and Geske, L., Speech Generating Device Use Among Children with Autism Spectrum Disorders a Caregiver Survey. 2005, University of Nebraska Medical Center, Maternal and Child Health Bureau, Health Resources and Services Administration, Department of Health and Human Services, Nebraska.

[81] Mirenda, P., Autism, augmentative communication, and assistive technology: what do we really know? Focus on Autism and Other Developmental Disabilities, 2001. 16: 231–243.

[82] Mirenda, P., Toward functional augmentative and alternative communication of students with autism: maneal signs, graphic symbols, and voice output communication aides. Language, Speech, and Hearing Services in School, 2003. 34: 203–216.

[83] Grimes, A., V. Kantroo, and R.E. Grinter, Let's Play!: Mobile Health Games for Adults, in Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, September 26–29, 2010. ACM, New York. p. 241–250.

[84] Hernandez, H.A., Z. Ye, T.C.N. Graham, D. Fehlings, and L. Switzer, Designing Action-Based Exergames for Children with Cerebral Palsy, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, April 27–May 2, 2013. ACM, New York. p. 1261–1270.

[85] Smeddinck, J., S. Siegel, and M. Herrlich, Adaptive Difficulty in Exergames for Parkinson's Disease Patients, in Proceedings of Graphics Interface 2013. Presented at the Graphics Interface (GI) Conference, Regina, SK, May 29–31, 2013, Canadian Information Processing Society, Toronto, ON. p. 141–148.

[86] Caro, K. Exergames for Children with Motor Skills Problems, in Proceedings of the 15th ACM SIGACCESS International Conference on Computers and Accessibility (ASSETS 2013), Bellevue, WA, October 21–23, 2013. ACM, New York. p. 1–8.

[87] Caro, K., F.L. Cibrian, L. Escobedo, C. Ramirez, A.I. Martínez-García, and M. Tentori, Diseño de un videojuego serio basado en movimiento para niños con el Trastorno del Desarrollo de la Coordinación, in Proceedings of the 2013 Mexican International Conference on Computer Science, Morelia, Michoacan, October 30 to November 1, 2013, C. Taller de Computación Clínica e Informática Médica, Editor.

[88] Caro, K., F.L. Cibrian, L. Escobedo, C. Ramírez, A.I. Martínez-García, and M. Tentori. Froggy Bobby: An Exergame for Children with Motor Skills Problems, in Workshop Ubiquitous Games and Gamification for Promoting Behavior Change and Wellbeing, Trento, Italy, September 16, 2013.

[89] Linney, C., A. Martinez-Garcia, and M. Tentori. Hunting Relics: A Collaborative Exergame on an Interactive Floor for Children. in Proceedings of the 2014 ACM International Joint Conference/Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct), Seattle, WA, September 13–17, 2014. ACM, New York. p. 223–226.

[90] Hsieh, G., S.A. Munson, M.C. Kaptein, H. Oinas-Kukkonen, and O. Nov, Personalizing Behavior Change Technologies, in Proceedings of the CHI '14 Extended Abstracts on Human Factors in Computing Systems, Toronto, ON, April 26–May 1, 2014. ACM, New York. p. 107–110.

[91] Nakajima, T., V. Lehdonvirta, E. Tokunaga, and H. Kimura, Reflecting Human Behavior to Motivate Desirable Lifestyle, in Proceedings of the 7th ACM Conference on Designing Interactive Systems, Cape Town, South Africa, 2008. ACM, New York. p. 405–414.

[92] Orji, R., J. Vassileva, and R.L. Mandryk, LunchTime: a slow-casual game for long-term dietary behavior change. Personal and Ubiquitous Computing, 2013. 17(6): 1211–1221.

[93] Tanenbaum, J.G., A.N. Antle, and J. Robinson, Three perspectives on behavior change for serious games, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, April 27–May 2, 2013. ACM, New York. p. 3389–3392.

[94] Billard, A., Robins, B., Nadel, J., and Dautenhahn, K., Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. Assistive Technology: The Official Journal of RESNA, 2006. 19: 37–49.

[95] Funk, M., R. Boldt, B. Pfleging, M. Pfeiffer, N. Henze, and A. Schmidt, Representing Indoor Location of Objects on Wearable Computers with Head-Mounted Displays, in Proceedings of the 5th Augmented Human International Conference, Kobe, Japan, March 7–8, 2014, ACM, New York. p. 1–4.

[96] Quek, F. and F. Oliveira, Enabling the blind to see gestures. ACM Transactions on Computer-Human Interaction, 2013. 20(1): 1–32.

[97] Sanchez, J., M. Saenz, and J.M. Garrido, Usability of a multimodal video game to improve navigation skills for blind children. ACM Transactions on Accessible Computing, 2010. 3(2): 1–29.

[98] Starner, T., Wearable Computing: Through the Looking Glass, in Proceedings of the 2013 International Symposium on Wearable Computers, Zurich, Switzerland, September 8–12, 2013. ACM, New York. p. 125–126.

[99] Kane, S.K., J.O. Wobbrock, and R.E. Ladner, Usable Gestures for Blind People: Understanding Preference and Performance, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, May 7–12, 2011. ACM, New York. p. 413–422.

[100] Lutz, R.J., Prototyping and Evaluation of Landcons: Auditory Objects that Support Wayfinding for Blind Travelers. Doctoral Consortium of the Eighth International ACM Conference on Computers and Accessibility (ASSETS '06). ACM SIGACCESS Accessibility and Computing, Portland, OR, October 23–25, 2006. ACM, New York. p. 8–11.

[101] Mehigan, T.J. and I. Pitt, Harnessing Wireless Technologies for Campus Navigation by Blind Students and Visitors, in Proceedings of the 13th International Conference on Computers Helping People with Special Needs: Volume Part II, Linz, Austria, July 11–13, 2012. Springer-Verlag, Berlin. p. 67–74.

[102] Rafael, I., L. Duarte, L. Carriço, and T. Guerreiro, Towards Ubiquitous Awareness Tools for Blind People, in Proceedings of the 27th International BCS Human Computer Interaction Conference, Lisbon, Portugal, September 9–13, 2013. British Computer Society, Swinton. p. 1–5.

[103] Castro, L., J. Favela, E. Quintana, and M. Perez, Behavioral data gathering for assessing functional status and health in older adults using mobile phones. Personal and Ubiquitous Computing, 2015. 19(2): 379–391.

[104] Escobedo, L., M. Tentori, E. Quintana, J. Favela, and D. Garcia-Rosas, Using augmented reality to help children with autism stay focused. IEEE Pervasive Computing, 2014. 13(1): 38–46.

[105] Caro, K., M. Tentori, A.I. Martínez-García, and I. Zavala, FroggyBobby: an exergame to support children with motor problems practicing motor coordination exercises during therapeutic interventions. Computers in Human Behavior, 2015.

[106] Alemdar, H. and C. Ersoy, Wireless sensor networks for healthcare: a survey. Computer Networks, 2010. 54(15): 2688–2710.

[107] Darwish, A. and A.E. Hassanien, Wearable and implantable wireless sensor network solutions for healthcare monitoring. Sensors, 2011. 11(6): 5561–5595.

[108] Hashemi, J., T.V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro. A Computer Vision Approach for the Assessment of Autism-Related Behavioral Markers, in Proceedings of the IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), San Diego, CA, November 7–9, 2012. IEEE, San Diego, p. 1–7.

[109] Mihailidis, A., G. Fernie, and J.C. Barbenel, The use of artificial intelligence in the design of an intelligent cognitive orthotic for people with dementia. Assistive Technology, 2001. 13: 23–39.

[110] De Silva, B., A. Natarajan, M. Motani, and K.-C. Chua. A Real-Time Exercise Feedback Utility with Body Sensor Networks, in Proceedings of the 5th International Summer School and Symposium on Medical Devices and Biosensors (ISSS-MDBS), Hong Kong, June 1–3, 2008. IEEE. p. 49–52.

[111] Waluyo, A.B., I. Pek, W.-S. Yeoh, T.S. Kok, and X. Chen. Footpaths: Fusion of Mobile Outdoor Personal Advisor for Walking Route and Health Fitness, in Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, September 3–6, 2009. IEEE, Minneapolis. p. 5155–5158.

[112] Mitra, S. and T. Acharya, Gesture recognition: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2007. 37(3): 311–324.

[113] Baer, D.M., M.M. Wolf, and T.R. Risley, Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968. 1(1): 91–97.

[114] McConnell, M.E., C.J. Cox, D.D. Thomas, and P.B. Hilvitz, Functional Behavioral Assessment: A Systematic Process for Assessment and Intervention in General and Special Education Classrooms. 2001: Love Publishing Company, Denver, CO.

[115] Peacocke, R.D. and D.H. Graf, An introduction to speech and speaker recognition. Computer, 1990. 23(8): 26–33.

[116] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, A Practical Guide to Support Vector Classification. 2003. http://www.csie.ntu.edu.tw/~cjlin/ (accessed January 8, 2016).

[117] Quinlan, J.R., Induction of decision trees. Machine Learning, 1986. 1(1): 81–106.

[118] Rabiner, L. and B.-H. Juang, An introduction to hidden Markov models. IEEE ASSP Magazine, 1986. 3(1): 4–16.

[119] Rodríguez, M.D., R. Martínez, M. Pérez, L.A. Castro, and J. Favela. Using Ontologies to Reduce User Intervention to Deploy Sensing Campaigns with the InCense Toolkit, in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, September 5–8, 2012. ACM, New York. p. 741–744.

[120] Bian, X., G.D. Abowd, and J.M. Rehg, Using Sound Source Localization to Monitor and Infer Activities in the Home. GVU Technical Report;GIT-GVU-04-20. 2004, Georgia Institute of Technology.

[121] Acampora, G., D.J. Cook, P. Rashidi, and A.V. Vasilakos, A survey on ambient intelligence in healthcare. Proceedings of the IEEE, 2013. 101(12): 2470–2494.

[122] Bryson, S.E., L. Zwaigenbaum, J. Brian, W. Roberts, P. Szatmari, V. Rombough, and C. McDermott, A prospective case series of high-risk infants who developed autism. Journal of Autism and Developmental Disorders, 2007. 37(1): 12–24.

[123] Salovey, P. and J.D. Mayer, Emotional intelligence. Imagination, Cognition and Personality, 1989. 9(3): 185–211.

[124] Tao, J. and T. Tan, Affective computing: A review, in Affective computing and intelligent interaction Tao, J., T. Tan, and R. Picard (Eds.). 2005: Springer, Berlin. p. 981–995.

[125] Zhang, H., G. Lopez, M. Shuzo, Y. Omiya, S. Mistuyoshi, S.i. Warisawa, and I. Yamada, A Database of Japanese Emotional Signals Elicited by Real Experiences, in Pervasive Computing Paradigms for Mental Health: 4th International Symposium, MindCare 2014, Tokyo, Japan, May 8–9, 2014, Springer, Cham. p. 3–12.

[126] Lopatovska, I. and I. Arapakis, Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. Information Processing & Management, 2011. 47(4): 575–592.

[127] Faja, S. and G. Dawson, Early Intervention for Autism. In Handbook of Preschool Mental Health: Development, Disorders, and Treatment ed. J.L. Luby. 2006: Guilford Press, New York. p. 388–416.

[128] Emiliani, P.L. and C. Stephanidis, Universal access to ambient intelligence environments: opportunities and challenges for people with disabilities. IBM Systems Journal, 2005. 44(3): 605–619.

[129] Mihailidis, A., J. Boger, T. Craig, and J. Hoey, The COACH prompting system to assist older adults with dementia through handwashing: an efficacy study. BMC Geriatrics, 2008. 8: 28.

[130] Levinson, R., PEAT: The planning and execution assistant and training system. The Journal of Head Trauma Rehabilitation, 1997. 12(2): 769–775.

[131] Sundberg, M.L. and J.W. Partington, Teaching Language to Children with Autism and Other Developmental Disabilities. 1998. Behavior Analysts: Pleasant Hill, CA.

[132] Mankoff, J., G. Hayes and D. Kasnitz, Disability Studies as a Source of Critical Inquiry for the Field of Assistive Technology, in Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility, Orlando, FL, October 25–27, 2010. ACM, New York. p. 3–10.

[133] Lehman, J.F. Toward the Use of Speech and Natural Language Technology in Intervention for a Language-Disordered Population, in Proceedings of the Third

International ACM Conference Assistive Technologies, Marina Del Rey, CA, April 15–17, 1998. ACM, New York. p. 19–26.

[134] Fell, H., J. MacAuslan, J. Gong, C. Cress, and T. Salvo, visiBabble for Pre-speech Feedback, in Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems, Montréal, Québec, April 22–27, 2006, ACM, New York. p. 767–772.

[135] Hailpern, J., K. Karahalios, J. Halle, L. DeThorne, and M. Coletto, Visualizations: Speech, Language & Autistic Spectrum Disorder, in Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems, Florence, Italy, April 5–10, 2008. ACM, New York. p. 3591–3596.

[136] Tobii Dynavox Compass. 2008; Available from: http://www.dynavoxtech.com/ (accessed January 18, 2016).

[137] Bowker, A. and S. Sennott, Autism, AAC, and Proloquo2Go. In American Speech-Language-Hearing Association (ed.), Perspectives on Augmentative and Alternative Communication. 2009: The Pennsylvania State University, University Park, PA. p. 137–145.

[138] Parés, N., A. Carreras, J. Durany, J. Ferrer, P. Freixa, D. Gómez, and O. Kruglanski. Promotion of Creative Activity in Children with Severe Autism Through Visuals in an Interactive Multisensory Environment, in Proceedings of the 2005 Conference on Interaction Design and Children, IDC'05, Barcelona, Spain, June 9–12, 2005. ACM, New York. p. 110–116.

[139] McDonnell, J., J. Waters, H. Balinsky, R. Castle, F. Dickin, W. Wah Loh, and K. Shepherd, Memory spot: a labeling technology. IEEE Pervasive Computing 2010. 9(2): 11–17.

[140] Piper, A.M., N. Weibel, and J.D. Hollan, TAP & PLAY: An End-User Toolkit for Authoring Interactive Pen and Paper Language Activities, in Proceedings of the CHI 2012, ACM Conference on Human Factors in Computing Systems, Austin, TX, May 5–10, 2012. ACM: New York. p. 149–158.

[141] Richard, E., Richard, P.; Gaudin, G., Augmented Reality for Rehabilitation of Cognitive Disabled Children: A Preliminary Study, in Proceedings of the International Workshop on Virtual Rehabilitation, Venice, Italy, September 27–29, 2007. IEEE Xplore. p. 102–108.

[142] Quill, K., Instructional considerations for young children with autism: the rationale for visually cued instruction. Journal of Autism and Developmental Disorders, 1997. 27(6): 697–714.

[143] Ryan, C.S. and S. Hemmes, Post-training discrete-trial teaching performance by instructors of young children with autism in early intensive behavioral intervention. The Behavior Analyst Today, 2005. 6(1): 1–12.

[144] Mueller, F.F., D. Edge, F. Vetere, M.R. Gibbs, S. Agamanolis, B. Bongers, and J.G. Sheridan. Designing Sports: A Framework for Exertion Games, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, Canada, May 7–12, 2011. ACM, New York. p. 2651–2660.

[145] Sinclair, J., P. Hingston, and M. Masek, Considerations for the Design of Exergames, in Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia, Perth, Australia, December 1–4, 2007. ACM, New York. p. 289–295.

[146] Juul, J., A Casual Revolution: Reinventing Video Games and Their Players. 2012: The MIT Press, Cambridge.

[147] Cai, S.X. and A.S. Kornspan, The use of exergaming with developmentally disabled students. Strategies, 2012. 25(3): 15–18.

[148] Parés, N., P. Masri, G. Van Wolferen, and G. Creed, Achieving dialogue with children with severe autism in an adaptive multisensory interaction: the 'MEDIATE' project. IEEE Transactions on Visualization and Computer Graphics, 2005. 11(6): 734–743.

[149] Wyeth, P., J. Summerville, and B. Adkins. Stomp: An Interactive Platform for People with Intellectual Disabilities, in Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology, Lisbon, Portugal, November 8–11, 2011. ACM, New York. p. 1–8.

[150] Burke, J.W., M. McNeill, D. Charles, P. Morrow, J. Crosbie, and S. McDonough, Serious Games for Upper Limb Rehabilitation Following Stroke, in Conference in Games and Virtual Worlds for Serious Applications, Coventry, UK, March 23–24, 2009. IEEE Computer Society, Coventry, UK, p. 103–110.

[151] Tsai, T., H. Chang, G. Huang, and C. Chang, WaterBall: the exergaming design for rehabilitation of the elderly. Computer-Aided Design & Applications, 2012. 9(4): 481–489.

[152] Burke, J.W., M.D.J. McNeill, D.K. Charles, P.J. Morrow, J.H. Crosbie, and S.M. McDonough, Optimising engagement for stroke rehabilitation using serious games. The Visual Computer, 2009. 25(12): 1085–1099.

[153] Sugarman, H., A. Weisel-Eichler, A. Burstin, and R. Brown. Use of the Wii Fit System for the Treatment of Balance Problems in the Elderly: A Feasibility Study. Presented at the IEEE Virtual Rehabilitation International Conference, Haifa, Israel, June 29–July 2, 2009. IEEE, p. 111–116.

[154] Deutsch, J.E., M. Borbely, J. Filler, K. Huhn, and P. Guarrera-Bowlby, Use of a low-cost, commercially available gaming console (Wii) for rehabilitation of an adolescent with cerebral palsy. Physical Therapy, 2008. 88(10): 1196–1207.

[155] Dunne, A., S. Do-Lenh, G. O' Laighin, C. Shen, and P. Bonato, Upper Extremity Rehabilitation of Children with Cerebral Palsy Using Accelerometer Feedback on a Multitouch Display, in Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, August 31 to September 4, 2010. IEEE Computer Society. p. 1751–1754.

[156] Hernandez, H.A., T.C. Graham, D. Fehlings, L.C. Switzer, Z. Ye, Q. Bellay, M.A. Hamza, C. Savery, and T. Stach. Design of an Exergaming Station for Children with Cerebral Palsy, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), Austin, TX, May 5–10, 2012. ACM, New York. p. 2619–2628.

[157] Finkelstein, S.L., A. Nickel, and E.A. Suma. Astrojumper: Designing a Virtual Reality Exergame to Motivate Children with Autism to Exercise, in IEEE Virtual Reality Conference (VR), Waltham, MA, March 20–24, 2010. IEEE, Boston. p. 267–268.

[158] Grønbæk, K., O.S. Iversen, K.J. Kortbek, K.R. Nielsen, and L. Aagaard. IGameFloor: A Platform for Co-located Collaborative Games, in Proceedings of the International Conference on Advances in Computer Entertainment Technology, Salzburg, Austria, June 15–17, 2007. ACM, New York. p. 64–71.

[159] Landry, P., J. Minsky, M. Castañer, O. Camerino, R. Rodriguez-Arregui, E. Ormo, and N. Pares. Design Strategy to Stimulate a Diversity of Motor Skills for an Exergame

Addressed to Children, in Proceedings of the 12th International Conference on Interaction Design and Children, New York, NY, June 24–27, 2013. ACM, New York. p. 84–91.

[160] Wyeth, P., D. Johnson, and J. Ziviani. Activity, Motivation and Games for Young Children, in Proceedings of the 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death, Melbourne, VIC, September 30 to October 1, 2013. ACM, New York.

[161] Linney, C., A. Martinez-Garcia, and M. Tentori. FroggyBobby: an exergame to support children with motor problems practicing motor coordination exercises during therapeutic interventions, ACM EA in Ubicomp 2014. Seattle, WA. September 13–17, 2014.

[162] Chung, K., M. Shilman, C. Merril, and H. Ishii. OnObject: Gestural Play with Tagged Everyday Objects, in Adjunct Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST), New York, October 3–6, 2010. ACM, New York. p. 273–282.

[163] Ayres, A.J. and L.S. Tickle, Hyper-responsivity to touch and vestibular stimuli as a predictor of positive response to sensory integration procedures by autistic children. American Journal of Occupational Therapy, 1980. 34(6): 375–381.

[164] Case-Smith, J. and Miller, H., Occupational therapy with children with pervasive developmental disorders. American Journal of Occupational Therapy, 1999. 53(5): 506–513.

[165] Villafuerte, L., M. Markova, and S. Jordá, Acquisition of Social Abilities Through Musical Tangible User Interface: Children with Autism Spectrum Condition and the Reactable, in CHI '12 Extended Abstracts on Human Factors in Computing Systems, Austin, TX, May 5–10, 2012. ACM, New York. p. 745–760.

[166] Ringland, K.E., R. Zalapa, M. Neal, L. Escobedo, M. Tentori, and G.R. Hayes, SensoryPaint: A Multimodal Sensory Intervention for Children with Neurodevelopmental Disorders, in Ubicomp, Seattle, Washington. 2014ACM, New York, p. 873–884.

[167] Zalapa, R. and M. Tentori, Movement-Based and Tangible Interactions to Offer Body Awareness to Children with Autism, in UCAmi '13 7th Internation Conference on Ubiquitous Computing % Ambient Intelligence, December 2–6, 2013. Springer International, Carrillo. p. 127–134.

[168] Jordá, S., The Reactable: Tangible and Tabletop Music Performance, in Proceedings of the CHI '10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, GA, April 10–15, 2010. ACM, New York. p. 2989–2994.

[169] Jordá, S., G. Geiger, M. Alonso, and M. Kaltenbrunner, The reacTable: Exploring the Synergy between Live Music Performance and Tabletop Tangible Interfaces, in Proceedings of the TEI '07 1st International Conference on Tangible and Embedded Interaction, Baton Rouge, LA, Febraury 15–17, 2007. ACM, New York. p. 139–146.

[170] Eyeclick. EyeStep from eyeclick, http://www.eyeclick.com (accessed January 6, 2016).

[171] Farr, W., N. Yuill, and H. Raffle, Collaborative Benefits of a Tangible Interface for Autistic Children, in Proceedings of the Workshop on Tangibles for Children (CHI 2009), Boston, April 4, 2009, ACM, New York. p. 1–4.

[172] Anderson-Hanley, C., K. Tureck, and R.L. Schneiderman, Autism and exergaming: effects on repetitive behaviors and cognition. Psychology Research and Behavior Management, 2011. 4: 129–137.

[173] Zuckerman, O., S. Arida, and M. Resnick. Extending Tangible Interfaces for Education: Digital Montessori-Inspired Manipulatives, in Proceedings of the SIGCHI Conference

on Human Factors in Computing System, Portland, OR, April 2–7, 2005. ACM, New York. p. 859–868.

[174] Antle, A.N. The CTI Framework: Informing the Design of Tangible Systems for Children, in Proceedings of the 1st International Conference on Tangible and Embedded Interaction, Baton Rouge, LA, February 15–17, 2007. ACM, New York. p. 195–202.

[175] Bartoli, L., C. Corradi, F. Garzotto, and M. Valoriani, Exploring Motion-Based Touchless Games for Autistic Children's Learning, in Proceedings of the 12th International Conference on Interaction Design and Children, New York, June 24–27, 2013. ACM, New York. p. 102–111.

[176] Gonçalves, N., J.L. Rodrigues, S. Costa, and F. Soares. Preliminary Study on Determining Stereotypical Motor Movements, in Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), San Diego, CA, August 28 to September 1, 2012. IEEE. p. 1598–1601.

[177] Tetteroo, D., A. Shirzad, M.S. Pereira, M. Zwinderman, D. Le, and E. Barakova. Mimicking Expressiveness of Movements by Autistic Children in Game Play. in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), Amsterdam, the Netherlands, September 3–5, 2012. IEEE, Los Alamitos. p. 944–949.

[178] Akten, M. The Mega Super Awesome Visuals Company. 2009.

[179] Noterdaeme, M., H. Amorosa, K. Mildernberger, S. Sitter, and F. Minow, Evaluation of attention problems in children with autism and children with a specific language disorder. European Child & Adolescent Psychiatry, 2001. 10: 58–66.

[180] Baranek, G.T. Efficacy of sensory and motor interventions for children with autism. Journal of Autism and Developmental Disorders 2002, 32(5), 397–422.

[181] Dawson, G. and Watling, R. Interventions to facilitate auditory, visual, and motor integration in autism: a review of the evidence. Journal of Autism and Developmental Disorders 2000, 30(5), 415–421.

[182] Rubenstein, L.Z., C.M. Powers, and C.H. MacLean, Quality indicators for the management and prevention of falls and mobility problems in vulnerable elders. Annals of Internal Medicine, 2001. 135(56): 686–693.

[183] Morris, A., R. Donamukkala, A. Kapuria, A. Steinfeld, J. Matthews, J. DunbarJacobs, and S. Thrun, A Robotic Walker That Provides Guidance, in Proceedings of the IEEE International Conference on Robotics and Automation, Pittsburgh, PA, September 14–19, 2003. IEEE, Piscataway. p. 25–30.

[184] Nourbakhsh, I., J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto, An affective mobile robot with a full-time job. Artificial Intelligence, 1999. 114(1–2): 95–124.

[185] Helal, S., S.E. Moore, and B. Ramachandran. Drishti: An Integrated Navigation System for Visually Impaired and Disabled, in Proceedings of the 5th IEEE International Symposium on Wearable Computers, Zurich, Switzerland, October 8–9, 2001. IEEE Computer Society, Washington, DC. p. 180–181.

[186] Duncan, J., L. Jean-Camp, and W.R. Hazelwood, The Portal Monitor: A Privacy-Enhanced Event-Driven System for Elder Care, in Proceedings of the 4th International Conference on Persuasive Technology, Claremont, CA, April 26–29, 2009. ACM, New York, pp. 1–9.

[187] Baus, J., R. Wasinger, I. Aslan, A. Kruger, A. Maier, and T. Schwartz, Auditory Perceptible Landmarks in Mobile Navigation, in Proceedings of the 12th International

Conference on Intelligent User Interfaces, Honolulu, HI, January 28, 2007. ACM, New York. p. 302–304.

[188] Goodman, J., S.A. Brewster, and P. Gray, How can we best use landmarks to support older people in navigation? Behaviour and Information Technology, 2005. 24(1): 3–20.

[189] Pentland, A., Healthwear: medical technology becomes wearable. IEEE Computer, 2004. 37(5): 42–49.

[190] Narzt, W., G. Pomberger, A. Ferscha, D. Kolb, R. Müller, J. Wieghardt, H. Hörtner, and C. Lindinger, Augmented reality navigation systems. Universal Access in the Information Society, 2006. 4(3): 177–187.

[191] Scott-Young, S., Seeing the road ahead: GPS-augmented reality aids drivers. GPS World, 2003. 14(11): 22–29.

[192] Tonnis, M. and G. Klinker. Effective Control of a Car Driver's Attention for Visual and Acoustic Guidance Towards the Direction of Imminent Dangers, in Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality, Santa Barbara, October 22–25, 2006. IEEE Computer Society, Washington, DC.

[193] Tentori, M., G.R. Hayes, and M. Reddy, Pervasive computing for hospital, chronic, and preventive care. Foundations and Trends in Human-Computer Interaction, 2012. 5(1): 1–95.

7

WIRELESS POWER FOR IMPLANTABLE DEVICES: A TECHNICAL REVIEW

Nikita Ahuja[1], Mary Mehrnoosh Eshaghian-Wilner[1], Zhuochen Ge[1], Renjun Liu[1], Alekhya Sai Nuduru Pati[1], Kodiak Ravicz[1], Mike Schlesinger[1], Shu Han Wu[1], and Kai Xie[2]*

[1] *EE-Systems, University of Southern California, Los Angeles, CA, USA*

[2] *School of Aerospace Science and Technology, Xidian University, Xi'an, Shaanxi, China*

7.1 INTRODUCTION

Wireless power transfer has been a hot topic in the recent years due to the wide use of mobile devices and their short battery life. Battery life is one of the essential parameters of a mobile device. Industrial designers are trying their best to make their products lighter and thinner. However, the battery often restricts their ideas and innovations. For example, in the iPhone 6 Plus smartphone, the battery weighs about 43 g [1], which contributes to 25% of the total weight (172 g) of the entire device [2]. The chip, screen, and the operating system of the device are all improving at a more rapid pace than the battery. In the medical field, implantable medical devices are being increasingly implemented as they are made smaller and smaller. Currently, power supply is the bottleneck in the industry design for both mobile systems and implantable medical devices.

Even small improvements in battery life can stimulate fast and convenient charging in the industry design. One of the most intuitive ideas is that of wireless power transfer, in which we charge the battery wirelessly. Human intelligence took

*The authors are listed in alphabetical order.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

more than a hundred years to devise this practical and reliable solution, and yet there are still many challenges to overcome.

Some people may think that wireless charging can be as powerful as wired charging. They may assume that it can be adapted to many scenarios without restriction. However, they may be disappointed with the current solutions for wireless power transmission. Current approaches are mainly focused on convenient charging, which means that users may not need plug-in charging anymore. Users might need to put their phone on a specially made wireless charging plate to enable wireless charging. The advantages are obvious—wires are no longer necessary, and we can save time and energy by eliminating the plug-in and plug-out action. The only requirement is to keep the phone on its particularly designed plate. However, the disadvantages of the current solutions cannot be ignored. Wireless charging is much slower and less efficient than wired charging is, both because some of the power generated is lost to the environment and because the circuits used for wireless charging have more complex designs and consume more power. Additionally, charging plates and receiver modules are sold as expensive accessories and therefore make wireless charging less desirable for consumers.

For implantable devices in the medical arena, wireless power transfer solutions are similar to those of industry. They share the same theory base and have similar approaches. However, differences between charging a phone and a cardiac pacemaker are significant. The distance between the transmitter and receiver depends on the requirements and can vary greatly—in normal phone, wireless power transfer distance is less than 1 cm, while the cardiac pacemaker may be as far as 20 cm or more. The rate of charge is also different. The maximum rate of the phone is often 5 V/2 A, while the rate of the cardiac pacemaker can be as small as 5 V/100 mA. In addition to these differences, there are a multitude of other parameters which must be met to construct a medical device. Ultimately, while the basic circuit for mobile devices may be used for implantable medical devices, there are many adjustments to be made. These restrictions forced us to be familiar with not only current industry solutions in the custom market but also solutions designed for specific purposes in the medical field. For this reason, we will mention the basic idea and theory of wireless power transfer in this chapter, and also talk about current industry and implantable solutions in detail.

Although the current solution is limited, there is a bright future for wireless power transfer. There is an estimate that the number of mobile phones will exceed the human population by 2014 [3]. The number of implantable medical devices will also significantly increase in the future. If this technology can be improved, applications like nanorobots and pervasive computing in the medical area can flourish.

In this chapter, we will first talk about the history of wireless power transfer technology. Next, we will talk about the background of the idea and the contribution that humans have made in the past hundred years. In the third part, we will primarily introduce the base theory of wireless power transfer. In the fourth part, we will present an approach that can be used for implantable medical devices. The standards that are widely used today in the industry and custom market will be introduced in the fifth part. In the sixth part, we will give some examples and ideas from the current demos in the medical area.

7.2 HISTORY OF WIRELESS POWER

In this section, a brief history of wireless power is presented. The information is categorized in three periods, as shown next.

7.2.1 First Period—Equations and Experiment from Maxwell Hertz and Tesla

When James Clerk Maxwell first introduced his famous equations in *Treatise on Electricity and Magnetism* in 1873, he published that power can be transmitted wirelessly from one point to another point by electromagnetic waves [4]. Later, Heinrich Hertz used experiments to prove Maxwell's equations [5]. At that time, Maxwell and Hertz might not have imagined that their contributions to wireless power and information transmission would significantly affect humans' lives for more than a hundred years. Even Tesla's unsuccessful experiments later on during the period of 1899–1910 did not eliminate the desire of humans to enable wireless power transfer.

7.2.2 Second Period—From Post–World War II to 1990s, the Large-Power Microwave Approach

Human intelligence had a breakthrough after the difficulties of World War II [6]. The evolution of electronic devices and research on wireless electronics significantly impacted the wireless power transmission research. In October 1964, an internationally televised helicopter flight that transmitted power via microwave to Spencer Laboratory of the Raytheon Company demonstrated the practical use of an aircraft that can fly at a high altitude [7]. The demo also demonstrated the remarkable potential of the microwave approach to revolutionize power transfer.

The theory breakthrough let the wireless power transfer concept mature. Goubau and Schwering [8] published an article that greatly accelerated the progress of the wireless power transmission. The paper introduced that power can be transmitted over any distance with almost 100% efficiency if the sender and receiver antenna are designed according to the instructions. This broke the traditional understanding that power density fell off as the square of the distance [8]. Their research was proven by Degenford who got over 99% efficiency in transmission using the 4 m wavelength microwave [9].

In the space area, NASA's Marshall Space Flight Center was also interested in a space station in the low-earth orbit. They were interested in figuring out if there was a solution that can wirelessly send power to a central space station, which then sends the power to the daughter satellites. The work demonstrated the two-mile power transmission with the efficiency of the system DC to DC power at 26% in 1974 compared to the efficiency of 13% in 1969 [6].

The solar satellite concept also influenced the development of wireless power transmission. A 30 kW of received power over a distance of 1 mile at JPL gold stone facility in the Mojave Desert was introduced. The gradual improvement in the

computer-aided calculation on the recent element gave the opportunity to increase the efficiency. The efficiency increased from 55% in 1964 to 91% in 1976 [10].

In the 1980s, people were interested in microwave-powered aircraft; several demos were introduced. A Canadian microwave-powered stationary high-altitude relay platform (SHARP) airplane was developed and an impressive demo was given to Dr. W. C. Brown [6] in Ottawa, Canada, in October 1987. The plane consisted of a 14 ft wing and a rectenna (a kind of antenna especially for microwave use). The plane used battery power to climb to the altitude of about 100 m. It then received the microwave power and switched the battery off. The microwave sender tracked the aircraft automatically, and the ground controller controlled the plane.

Up until the late 1980s, most of the technology was focused on the bandwidth around the 2.4 GHz frequency, even though the higher frequencies closer to 35 GHz could be much more useful and efficient. However, it needed the support of the antenna design and amplifier to work at that frequency. In 1992, Peter Koert from Arco Power Technologies introduced a new approach to design a 35 GHz rectenna with a possible efficiency around 50–70% [11]. Nowadays, the systems work at 35 GHz frequency, which is mature and the demo of this system is successful. Before the twenty-first century, people were still trying to find an excellent solution for wireless power transmission in daily use. Thanks to a few generations of researchers' efforts the breakthrough came at our age [6].

7.2.3 Third Period—From Large to Small, Goes to Everyone's Life

Although many efforts have been made in the power transmission via microwave, the sender and receiver are not small enough for mobile and portable devices. On the other hand, improvements to solar energy and fuel cells put the wireless power transfer in an difficult situation. Wireless power transmission technology is being replaced by renewable energy sources in the industry area, while the customer market still needs a more portable, less expensive, and reliable solution. Although researchers tried in many ways [12–17], there was still a long way to go as of 2000. The key was found by Kurs from MIT in 2007 [18]. They took advantage of electromagnetic resonance to realize a midrange (2 m) wireless power transfer to light a 60 W bulb, with an efficiency of 40–50%.

Electromagnetic resonance transmission has an advantage compared to the previous method of electromagnetic induction and microwave transmission. Compared to the electromagnetic induction approach, electromagnetic resonance transfer can significantly enlarge the range of the transmission. The usual electromagnetic induction method can only transfer the power in less than 1 cm [12]. However, the electromagnetic resonance transmission range can be improved in a different environment. Compared to the microwave power transmission, the amplifier, sending or receiving antenna, and circuit design can be simpler and work in the lower frequency band. Also, this new approach has less electromagnetic radiation (EM) production [19], which would be less harmful to the human body and provide a wider potential for application in medical implant devices.

The introduction of this theory and the power chip triggered the market cannibalization in wireless power transmission in the mobile and portable device market. The market will exceed $11B by 2020 [20]. So, it is an excellent opportunity to extend this technology to the medical area. In this chapter, we will mainly discuss the electromagnetic approach that is also widely used in the industry and medical areas.

7.3 APPROACH OF WIRELESS POWER TRANSMISSION

In this section, we discuss the various methods of transferring power wirelessly. The approaches discussed are microwave, inductive coupling, and magnetic coupling resonance.

7.3.1 Microwave

The central idea of wireless power transmission is that the power is transferred to the transmitting antenna after converting electrical energy into microwave. Finally, the power is received by the receiving antenna, and the receiver circuit transfers the microwave into electrical energy. Figure 7.1 shows the block diagram of how microwave wireless transmission occurs.

The prominent characteristics of this approach are as follows:

1. System works in the microwave frequency so that the system design can be similar to the microwave communication system. There are lots of assorted devices and components that can be adapted.
2. In long distances, power lines are tensioned by the earth's gravity and require sufficient strength. However, wireless power transmissions with microwaves are not subject to such an interference. For this reason, NASA was interested in the space wireless power transmission in the 1970s. In Figure 7.2, the block diagram shows how microwave is propagated in space.
3. The power transmission may have sharp directivity by using a focus antenna or parabolic antenna, which means that it can realize the point to point power transmission. As the direction can be controlled by the antenna, the target receiver can switch quickly, which is hard for wired transmission.
4. Less energy is lost to the air. Theoretically, the microwave transmission can be less lossless than wired transmission [21]. The system can get high efficiency in practical usage.

FIGURE 7.1 Microwave wireless transmission.

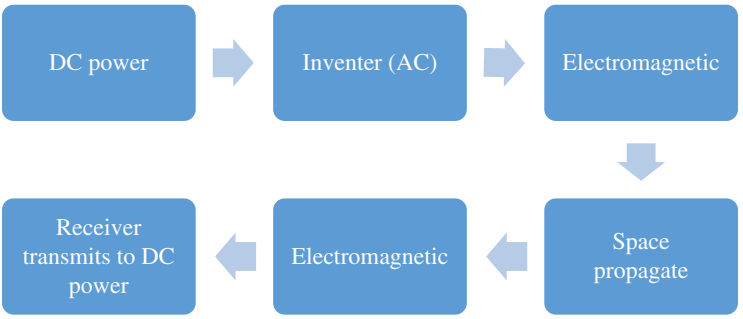


FIGURE 7.2 Microwave propagation in space.

7.3.2 Inductive Coupling

Power transmission via electromagnetic induction is based on the theory of Maxwell: a changing magnetic field can produce an electric current. However, this idea may not be practically used in the wireless power transmission because of the low efficiency of the transmission and the fact that coupling coefficient will decrease whenever distance changes from the ideal. As a result, it is not practical to transfer the power via electromagnetic induction.

In the 1980s, Yaskawa of the Japan National Research Institute proposed the approach of inductive power transfer. In the 1990s, researchers from the University of Auckland did more research in this area, and the technology had its official name: inductively coupled power transfer (ICPT). Prof. Boys and his research team issued a few patents in this field [22, 23].

First, the DC would be converted to the high-frequency AC power, then the coils that behave as the sender and receiver will have the electromagnetic induction effect. The power will be then transferred from the originator to the receiver [24]. There are a few successful applications that take an advantage of inductive coupling wireless power charge, such as the wireless rechargeable toothbrush [25]. As the electrical toothbrush always has to work and is stored in a humid environment, wireless charging can increase both the service life and the reliability of the device.

7.3.3 Magnetic Coupling Resonance

This technology is based on the electromagnetic resonance theory. The transmitter and receiver coils are configured with the same resonant frequency. The resonance increases the distance of wireless power transmission. In the same driving voltage or current, the resonance increases the intensity of the magnetic and electric fields by Q times, which in turn increases the transmission distance [26].

Maxwell's electromagnetic theory presented the correlation between the electric and magnetic fields. The continuous changing of the electric field will produce the magnetic field; similarly, the continuous changing of the magnetic field will produce the electric field. The dependency of these two areas is called the electromagnetic field. The electromagnetic field has two different ways to propagate: near field and

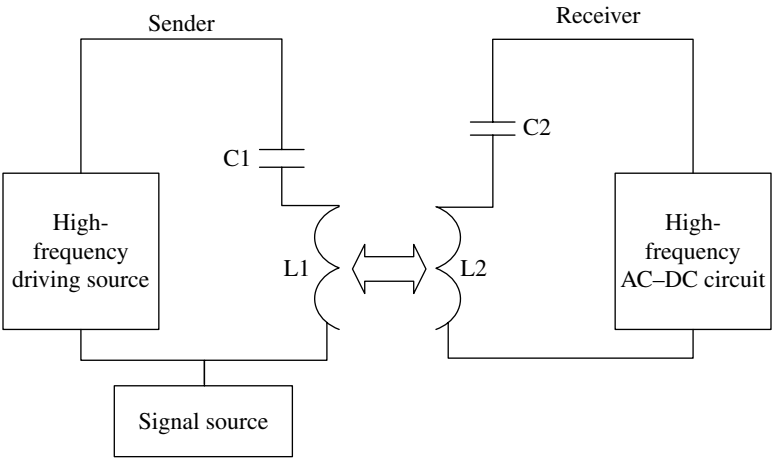


FIGURE 7.3 Magnetic resonance wireless power transmission system architecture.

far field. The spread in the near field is mainly by induction and the far field is primarily due to radiation [27].

The strength of the magnetic field is determined by the distance in the near field. Also, the power distribution of the magnetic field in the near field is not uniform. The energy quantum is always near the sender and moves around without the propagation of radiation. For this reason, if the receiver is in the near field, the changing magnetic field can be intermedium (a medium that plays intermediate between the power transfer system) to play the role of power transfer. In the system of magnetic coupling resonance wireless power transmission, both the sender and the receiver part include a resonator. When the sender is driven by the external signal whose frequency is close to the resonance frequency, the system will go to the resonance condition. During the resonance condition, the power will be transferred from the sender to the receiver. In the resonance transmission system, the power that has been sent but not received will be temporarily stored in the resonance circuit and magnetic field near the coil. It will be absorbed by the resonance circuit in the next half cycle. When the resonance frequency is low such that the system does not emit a far-field radiation pattern, this temporarily stored energy will not be lost. As a result, the efficiency will be increased [18].

Figure 7.3 shows the architecture of the magnetic resonance wireless power transmission system. The high-frequency driving source voltage is the sending DC voltage, C_1 and L_1 constitute the sender series resonance circuit, and L_2 and C_2 constitute the receiver series resonance circuit. The signal source has a frequency f. It offers the alternating electric field in the sender. From the resonance theory, we can conclude that when the frequency of the signal source, sender resonance, and receiver resonance is equal, the transmission efficiency can reach the peak. If the circuit works at its peak efficiency, it can qualify as operating under the electromagnetic resonance operating condition [28].

The power of the magnetic field when the resonance system is working is determined as follows:

$$\omega_{\mathrm{m}} = \frac{1}{2} LI^2 \tag{7.1}$$

Similarly, the power of the electronic field is determined as follows:

$$\omega_{\mathrm{e}} = \frac{1}{2} CV^2 \tag{7.2}$$

V is the voltage across the capacity and I is the current in the circuit.

When the resonance system is working, the electronic and magnetic fields will have power exchanges according to the period, which is determined by resonance frequency. The coil conserves the energy of the magnetic field, and the capacitor conserves the energy of the electric field. It is essential to choose the capacitor and coil in the system carefully in practical use. The efficiency of the resonance circuit is the most important part of the system. The high or low effectiveness directly determines the quality of the wireless power transmission system.

7.4 A DETAILED EXAMPLE OF MAGNETIC COUPLING RESONANCE

This section presents an approach that can be used for implantable medical devices. The design includes the four-coil model. The design concept and procedure can also be adopted for the other implantable devices. On the other hand, the model considers the balance between practicality, cost, size, and efficiency.

7.4.1 Theoretical Model

This section explains the basic parameters that are used in the resonance system (Fig. 7.4).

7.4.1.1 Inductance The self-inductance coefficient measures the closed current-carrying coil magnetic flux, and it has no relationship to the electronic circuit. The self-inductance coefficient is determined by the size, shape, number of windings,

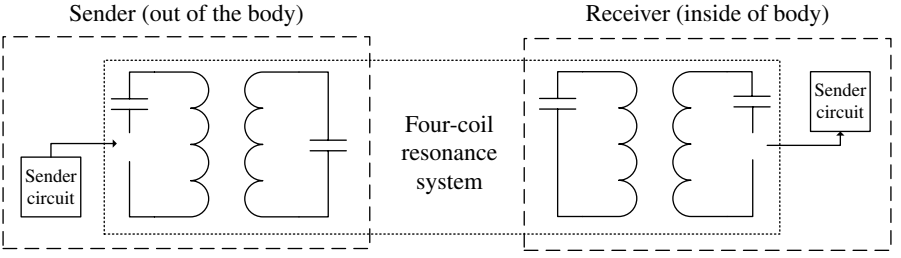


FIGURE 7.4 The-four coil resonance system.

and the surrounding (especially the inside of the internal of the coil) permeability of the magnetic medium. To realize a high-Q inductor coil, it is wise to use multilayer plane spiral and multiturn coils [29].

7.4.1.2 Parasitic Capacitance Generally speaking, the multiturn inductor will produce parasitic capacitance. The parasitic capacitance causes the self-resonance phenomenon that will limit the working frequency of an inductor. The tightly wound coils that do not have the space between neighbor turns will have less interference in the power transfer; as a result, it will be proper in practical use [18].

7.4.1.3 Alternating Current Resistance The alternating current resistance of an inductor resists the Q value of the inductor coil. It is essential to maximize the efficiency of the inductor coil to reduce the alternating current resistance. Strand wired Litz coil can effectively decrease the resistance of alternating current [18].

7.4.1.4 System Combined Model—Four-Coil Magnetic Resonance Transmission System The four-coil magnetic resonance power transfer consists of sender coil ($coil_1$), resonance $coil_1$ ($coil_2$), resonance $coil_2$ ($coil_3$), and receiver coil ($coil_4$). The four-coil model is based on the resonance circuit theory, and its power transmission equivalent circuit model is shown in (Fig. 7.5).

According to the magnetic resonance theory, if the working frequency is equal to the self-resonance frequency, the system transmission efficiency can reach the maximum [30]. At that time, the transfer efficiency is a function of the send and receive coils' Q values and two coils' coupling coefficient.

7.4.2 Improvements in the Four-Coil Resonance Model

In this section, we show the improvements that can be made to the four-coil system for improving transfer efficiency. We have presented four improvements. We also discuss the issues faced while implanting the four-coil resonance system into the body and how those issues are resolved.

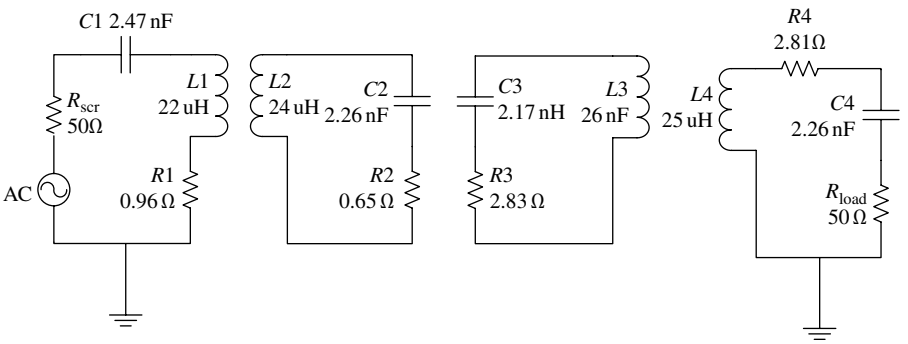


FIGURE 7.5 Four-coil method.

7.4.2.1 Winding Resonant Inductor The radio frequency at 100 kHz to 4 MHz is less harmful to the human body than microwave and higher frequency radiation. The human body absorbs less radio frequency signal when the frequency is low. Using lower frequency is a good choice for the implantable medical device. The varnished wire is widely used in the industry area and is made of inductor. However, the skin effect and the bulk effect will increase the AC resistance, which gives the big Q value inductor coil [31].

One of the approaches is using multistrand Litz coil. We can use sixty AWG44 (a kind of dimension parameter) varnished wires combined as a sender coil. Also due to the limited space for the receiver coil, we can use seven AWG44 varnished wires as a receiver coil. At the working frequency (around 350–850 kHz), AWG44 varnished wires of Litz coil keep the balance between the large Q value, lower AC resistance, and smaller size.

7.4.2.2 Inductor Model Size Design The model and size of an inductor design is mainly focused on the inductor coil layer number and the number of turns in each layer. These two facts affect the Q value of the inductor in practical use. The inductance value will increase by the number of turns. We designed a simple experiment using the varnished wires to make an inductor.

We use 85 mm coaxial plastic pipe to wind the inductor using the varnished wires. The diameter of varnished wires is 0.3 mm, and the diameter of the coil is 85 mm. Table 7.1 shows that the more the winding of varnished wires, the bigger the inductance value we can acquire. We recommend 10 turns for coil_1 and coil_4 and 30 turns for coil_2 and coil_3 that will have higher Q value and increase the transmission efficiency.

We can conclude that the more the layer (N_a) of the inductor, the more the self-inductance value. At the same time, with the increasing number of turns, the self-inductance value will increase. However, the growing number of turns and growing number of the layers will also increase the AC resistance. Also, the increasing working frequency will also increase the inductor AC resistance, which will decrease the Q value of the inductor. The balance between these factors should be determined by the design requirement in practical use (Fig. 7.6).

7.4.2.3 Resonance Frequency Design In this equation, L_a is the coil's self-inductance, and C_{self} is the parasitic capacitance. According to the previous conclusion, coil_1 and coil_4 have the inductance of 16 uH, and coil_2 and coil_3 have

TABLE 7.1 Effect of Turn Number on Inductance

Turn Number	Diameter (mm)	Testing Frequency (MHz)	Inductance Value (uH)
8	85	1	9
10			16
20			30
30			56

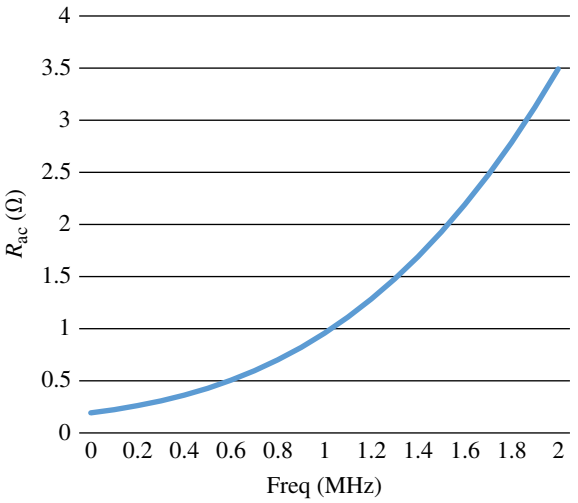


FIGURE 7.6 The relationship between AC resistance and operation frequency.

TABLE 7.2 Design Parameters of Enameled Wire Inductor

Coil Number	Diameter of Coil (mm)	Tune Number	Inductance (uH)	Series Capacitance
$Coil_1$	85	10	16	2.20
$Coil_2$		30	56	0.56
$Coil_3$		30	56	0.56
$Coil_4$		10	16	2.20

the inductance of 56 uH at the frequency of 1 MHz, respectively. Single-coil self-resonance frequency can be determined as follows:

$$f_{\text{self}} = \frac{1}{2\pi\sqrt{L_a C_{\text{self}}}} \tag{7.3}$$

In our design, the four coils should work in the same resonance frequency. To do so, we should also consider the value of the capacitance. The result is that $coil_1$ and $coil_4$ should have the capacitance of 1.56 nF and $coil_2$ and $coil_3$ should have the capacitance of 0.45 nF, respectively. However, the limit for the capacitance value of $coil_1$ and $coil_4$ is 2.2 nF. $Coil_2$ and $coil_3$ are limited at 0.56 nF (Table 7.2).

We also set the working frequency as 650 kHz and load resistance $R_{\text{load}} = 50\,\Omega$.

After the calculation of the previous model and equations, we can get the optimized solution of the coil model (Tables 7.3 and 7.4).

7.4.2.4 Human Body and Skin Effect The receiver of the four-coil transmission system might be surrounded by the human body and muscular tissue, and so we should also take human tissue effect into consideration [32]. We first consider the

TABLE 7.3 Dimension of Each Coil

Coil Number	Varnished Wire Diameter	Number of Varnished Wires in Every Tune	Layer Number	Tune Number	Inner Diameter of the Coil	External Diameter of the Coil
$Coil_1$	0.05	60	2	10	34	36
$Coil_2$	0.05	60	2	10	36	38
$Coil_3$	0.05	7	5	6	16	18
$Coil_4$	0.05	7	5	6	14	16

TABLE 7.4 Optimum Design of Coil Inductor Model Based on Theory Model

Coil Number	External Diameter of the Coil	Inner Diameter of the Coil	Tune Number	Layer Number	L (uH)	Q (Loaded)
$Coil_1$	36	34	10	2	25	1.8
$Coil_2$	38	36	10	2	26.5	75
$Coil_3$	16.5	14	6	5	26.4	29
$Coil_4$	19	16.5	6	5	24.8	2

muscular tissue effect to the implantable coil self-inductance (L_{eff}), parasitic capacitance (C_{self}), AC resistance (R_{ac}), self-resonance frequency (F_{self}), and Q value. The human muscular tissue consists of a few free magnetic materials, so the permeability is close to the vacuum environment. All the coil inductance is determined by its structure, size, and surrounding permeability. As a result, coil implanted in the human body has nearly the same inductance to the coil in the usual environment. However, the muscular tissue has a very high dielectric constant, compared to the regular environment, and hence the capacitance of the coil implanted in the human body will increase. Coil AC resistance is determined by the permeability and has no relationship with dielectric constant, so the AC resistance of an implantable device in the human body will almost remain the same. As mentioned earlier, the parasitic capacitance will decrease in the human body so that the self-resonance frequency will decrease and finally the Q value also decreases.

Although the effect may seem considerable, we can reduce the self-resonance frequency even more to increase the efficiency. In addition, due to the system's lower work frequency (which is less than 4 MHz), the implantable coil capacitance will be far bigger than the parasitic capacitance. As a result, the human body will not affect the working of the system to some extent.

The last point is that it will produce little electric current when the magnetic field passes through the human body. However, the low radio-frequency signal will be absorbed less than the high-frequency signal by the muscle. So, another advantage for the system designed with lower work frequency is that the human body will have less interference with the power transmission.

Using all the parameters mentioned before, we can realize the improved system. The system's working frequency is set at $f_{\text{self}} = 650\,\text{kHz}$. The coupling efficiency of

the coils is the function of the distance between two coils. Also, we can get the theoretical efficiency about 70% when $d = 10\,\text{mm}$, $f = 650\,\text{kHz}$, and k_{23} is in good condition [32].

7.4.3 Result and Test Method

In this section, we list the different tests used to check the functionality and efficiency of the four-coil resonance system. We briefly describe the test to check if the resonance system emits lower than harmful radiation for the human body. We also briefly describe a test to check the safety of implanting the device inside the human body.

7.4.3.1 Primary Test Essential test items include the sender input voltage and current as well as receiver output voltage and current. It is a basic test where the functionality of the device is tested by giving a few inputs and checking for the required outputs.

7.4.3.2 Electromagnetic Compatibility The wireless power transmission will emit radiation because of the changing magnetic field and electric field. Although the frequency is low, the switching power supply design of the sender and receiver may also emit a considerable amount of radiation, which will do harm to the human body. We should not only consider the working condition but also test the nonloaded circumstances. While doing the test and design, we can take advantage of filtering, shield, the ground connection, and printed circuit board (PCB) design to minimize the radiation emission.

7.4.3.3 Safety Test We should also make sure of the safety of the device. The most important issue of the wireless power transmission devices is the increasing temperature while working. The efficiency of wireless power is lower than the wired charge. As a result, the system will release more heat. The high temperature will not only destroy the circuit but also do harm to the human body, which has the implantable device.

7.5 POPULAR STANDARDS

Sections 7.3 and 7.4 analyze the fundamental theory of the wireless power transmission method and give a model that explains in detail how to design the circuit for an implantable device. However, there are still lots of mature wireless power transmission solutions that we can adapt. There are three main standards, and each of them have different application scenarios, advantages, and disadvantages.

7.5.1 Qi Standard

Qi is introduced by the Wireless Power Consortium (WPC), which has the feature of portability and universality. WPC is the first organization worldwide to formulate the standard for wireless power transmission and the largest organization in the world.

The members are from 15 different countries and 137 cooperative partners including HTC, Nokia, LG, Motorola, and Samsung. There are more than a hundred devices that support Qi charge. Qi standard is based on the electromagnetic induction wireless power transmission concept. There is one coil at both the receiver and the sender end [33, 34].

The Qi standard advocates the small coil design, which can transfer power in higher frequency, and the size of the sender and receiver can be small and flat. However, the weakness of the Qi standard is also evident; the charge and transfer distance is limited. The maximum distance is less than a few centimeters [35]. The typical design of the sender is a flat charger; the user should put their device on the flat charger to enable the charging. Another weaknesses of the Qi standard is that it only supports single device charging at one time [36]. To conquer the obstacles, intelligent engineers presented a solution to install more than one coil in the sender flat [37]. Although it sounds reasonable and smart, undoubtedly it will have more power wastage and will increase the cost of manufacturing. Also, users should always place the device accurately at the inducing field of the flat to ensure the maximum efficiency of the charging [38].

The Qi standard has a communication mechanism to reduce the power consumption during the idle time. Also, the communication mechanism between the receiver and sender can exchange the charging information like the charging speed, battery condition, etc. to make the charging more safe and efficient. The sender unit will adjust the different modes to maximize the load speed or minimize the power consumption [39]. One of the unavoidable weakness of the Qi standard is that the electromagnetic induction may heat the conducting material in the device. This weakness will cause high risk of fire or reduce the lifetime of the device if the sender is not designed properly [36].

Qi can support less than 5 W power consumption device charging nowadays, for example, smartphone, pad, and digital camera. The ideal efficiency is around 50% that is lower than the efficiency of the wired charger [40]. However, Qi support device has strived in the recent years. It shows a high potential opportunity in the future. In the medical device area, the Qi standard can be used as the portable device charger and a particular device with a closed shell. However, we think that due to the limitation of the range, the Qi standard might not be suitable for the implantable devices in the human body.

7.5.2 Power Matters Alliance Standard

Power Matters Alliance (PMA) standard is introduced by Duracell Powermat, which is led by Powermat and P&G. The concept of PMA standard supports the IEEE standard phone or electronic devices. PMA now has partners like ZTE, Blackberry, Starbucks, AT&T, Google, and other companies [41]. PMA standard takes advantage of two ways to charge the battery—first is the chip and PCB board inside the device, and the second is charging by wireless charging card (WiCC). The card is easy to use, light, and portable. The only thing the user should do is to stick the charging card on the battery [42]. On the other hand, WiCC can be used as

near-field communication (NFC) antenna [43]. It seems that PMA standard has a bright future. However, the pitfalls of PMA standard cannot be ignored. The non-changeable device like iPhone cannot adapt to WiCC. In order to enable the function, the manufacturer of the instrument should add the connecting terminals of the battery—two for the battery charge as normally required, two for the NFC signal transfer, and two for the data communication. Nowadays, the devices in Starbucks have adapted the PMA standard [44]. We can imagine the future scenario where we can charge the device during drinking coffee. PMA standard is similar to the Qi standard, and we think it is a good idea for a portable device. However, due to the same theoretical concepts as the Qi standard, the transfer range might also be limited, which will be the obstacle to the wireless power transmission for implantable devices.

7.5.3 Alliance for Wireless Power Standard

Alliance for Wireless Power (A4WP) is introduced by Qualcomm Inc., Samsung, and Powermat. The goal of A4WP is to make the standard for all the electronic devices, even the electric-powered car. A4WP takes advantage of the electromagnetic resonance technology that is different from the other two standards [45]. Compared to the Qi standard, A4WP has a larger coil at both the sender and the receiver end [36]. Also, this standard supports the charging for more than one device at the same time. Also, the accuracy setting for the resonance frequency enables more efficient charge and can support larger distances. A4WP can also support the nonaccuracy position charging, which means that the user can use devices with more freedom [45]. Similar to the Qi standard, A4WP can support the adaptive charging strategy. Also, due to the different approach from the Qi standard. A4WP can support more devices and more application scenarios like make the charging fault smaller and different charge condition [36].

There are a few devices that support A4WP charging. Also, the efficiency is a challenge in practical use [36]. A4WP standard will also have the challenge of Qi and PMA standard. In the medical area, we think that not only the portable devices but also the implantable devices may be suitable for A4WP charging. A4WP can support more devices and has a more advanced design concept. However, due to the late introduction and fewer manufacturers, A4WP may cost more and be more complicated to use.

7.6 WIRELESS POWER TRANSMISSION IN MEDICAL USE

In this section, we list a few functions of implantable devices in the field of medicine and their advantages. We discuss the approaches and solutions to the power transmission in implantable medical devices. We also discuss the compatibility and safety of implantable devices in humans. We also compare the two approaches used by transcutaneous energy transmission system (TETS).

TABLE 7.5 Classification of Implantable Electronic Devices

Class	Name	Function
Implantable stimulator	Pacemaker	Make heart beat by electric current
	Brain pacemaker	Wake up vegetative state and treat depression
	Vibration control	Stimulate thalamic and deal with Parkinson's disease
Implantable measuring	Fixed	Measuring physiological and biochemistry parameters
	Capsule type	Diagnosing the digestive tract
Artificial organ	Heart	Repair or replace the cardiac structure
	Brain	Simulate the human brain
	Cochlea	For hearing rehabilitation
Implantable drug delivery		Give drug directly

7.6.1 Introduction to Implantable Medical Devices

Implantable medical devices have many functions, which help in replicating the functions of the human organs (Table 7.5). The following classification of implantable devices can be made based on function:

1. Implantable stimulator, like cardiac pacemaker, defibrillator
2. Implantable measuring system, like capsule endoscopy
3. Implantable artificial organs, like artificial heart
4. Implantable medical devices, like drug pump

Apart from the functions of these devices, they have many advantages. We can obtain the test data continuously from day to night. Test data can be obtained without going through the skin, and that can reduce the interference from the skin. We can assist in curing the diseases that external devices cannot, like Parkinson. We can replace the function of normal organs like the heart, retina, and cochlea. We get a direct contact with organs, which is the most efficient approach.

7.6.2 Approaches and Solutions for Power Supply

The implantable device, unlike other typical external devices, has a different design requirement. The implantable device should work reliably and have a long duration. It is essential to adopt a proper solution for power supply (Table 7.6). The power supply should be small, have a low self-discharge rate, proper sealing, and biocompatibility [46]. Different devices have a different power supply requirement [46].

In the early years, wired power transmission was the most common and natural way. For example, the artificial heart Cardiowest is powered by wired supply. The wire goes through the human skin and directly connects with Cardiowest [47]. In this system, the human body will have potential infection risk.

Nowadays, there are two main approaches to the power supply:

TABLE 7.6 Power Consumption of Different Devices

Name	Power Requirement
Pacemaker	30–100 uW
Artificial cochlea	100 mW
Implantable drug delivery	2 mW
Neuromuscular stimulator	30 uW
Capsule type	50 mW
Artificial heart	12–45 W

7.6.2.1 Implantable Battery Evolution Batteries designed for implantable devices used to be nonreplaceable and nonrechargeable. The batteries in the early years included a nickel–cadmium battery (Ni–Cd) and Zinc–Mercury cell. However, these two solutions are not perfect enough compared to the lithium battery. The lithium battery-powered cardiac pacemaker has 4.7 years battery life [48]. There are more than five million lithium battery-powered cardiac pacemakers that have been put into daily use. Although lithium battery is an advanced solution, it still needs to be changed by surgery.

7.6.2.2 Transcutaneous Energy Transmission System TETS is similar to wireless power transmission. Consider the different working frequency and working power, there are two separate solutions:

Radio-Frequency Power Transmission This concept is analogous to the microwave wireless power transmission approach. In 1975, Nathan O. Soda introduced a simple and useful Class-E amplifier that took advantage of MHz work frequency to realize wireless power transmission. This method has been improved for more than 40 years, and it is an appropriate solution when the power-consuming device is small and has low-power dissipation. Today, artificial cochlear and the retinal prosthesis are powered by radio-frequency power transmission [49–51].

Contactless Transformer Approach The increasing power consumption and the complexity of the electronic system are like the artificial heart that cannot be satisfied by the radio-frequency power transmission approach. The contactless transformer approach has been widely adopted [52–54]. The design theory is as similar as electromagnetic resonance model [52–55].

7.6.2.3 Other New Power Supply Approaches The improvement of the new source power gave an idea to the intelligent researchers about using magnetic, sound, light, and heat energies to provide the power of implantable devices [56–59]. One of the interesting applications is using piezoelectric materials that can convert the shake (vibrant into electricity) pressure into electricity to supply the implantable device. In 1995, Williams and Yates introduced a device powered by environmental disturbance [60]. Glynne-Jones improved this method to have the output power at 3 uW by the 80.1 Hz environment noise [61]. This idea can also be used in the implantable devices.

The mechanical power that is produced by a human walking can also be a harvest and convert into electricity. J. Kymissis introduced an approach to collect the impact energy from the foot while people are walking. The implantable devices can obtain the power supply from this device [59]. In the experiment, the load at 10 Ω can measure 0.23 W of electricity power.

Another interesting approach is using infrared radiation to supply the power for a cardiac pacemaker [62]. The design enables an efficiency of 19% power transmission and requires a 17 min charge to realize 24 h of battery life.

7.6.3 Biocompatibility and Human Safety

A research team from the University of Pennsylvania tested the health effects of the TETS-powered ventriculus sinister assistant device [63]. They implanted a 160 kHz, max power 16.9 W TETS system into a cow's body. The maximum living time was 174 days among the 12 cases, and there was no biocompatibility issue that caused the termination of the experiment.

Kenji Shiba, a researcher from Hiroshima University, was doing research on the TETS-specific absorption rate (SAR) [64]. The test used a 20 W TETS system as the study object. The working frequency was from 200 kHz to 1 MHz. The conclusion was that SAR increases with the working frequency. When the frequency comes to 1 MHz, the SAR is 14 mW/kg. This is much lower than the International Commission on Non-Ionizing Radiation Protection (ICNIRP) standard limit of 2 W/kg.

7.6.4 Comparison of the Two Different TETS Technologies

The Class-E circuit is simple, small, and easy to integrate, and it is widely used in some small applications, especially suitable for the required size limit, and high-integration application scenario. However, Class-E structure can send limited radio-frequency power. G. X. Wang from the University of California successfully demonstrated the retinal prosthesis powered by Class-E system. It can transfer 250 mW power at the distance of 15 mm, frequency rate of 1 MHz, coils coupling coefficient of 0.08, and efficiency of 38.5% [65].

The resonance TETS system is suitable for the compound and large power consumption scenario (Table 7.7). Masaya Watada research focuses on the TETS system for the ventriculus sinister (a part of the human body, which is near to the heart) where the working rate is 60–120 kHz, and it has the efficiency of 40% with the transmission power 8 W and distance 5 mm [66].

We can conclude that Class-E is easier, costs less, and is more suitable than other solutions for the low-power consumption implantable device. TETS is more powerful and has a higher efficiency, and it is appropriate for the compound system.

7.7 CONCLUSION

Wireless power transmission can be widely applied in both the industrial and the medical areas. This technology offers an alternate approach to wired power supplies. Although the history of this technology is spread over the course of one hundred years,

TABLE 7.7 E Type and Full-Bridge Type TETS Technology with the Comparison of Different Parameters

Circuit Topology	E (Parameter Set 1)	E (Parameter Set 2)	Full-Bridge (Parameter Set 1)	Full-Bridge (Parameter Set 2)
Winding interval (mm)	1–12	1–10	0.8–14.4	10–20
Structure	OD 30 mm	OD 25 mm	OD 40 mm	OD 66 mm
Coupling coefficient	0.10–0.46	0.10–0.40	0.17–0.85	0.16–0.39
Power	7 mW	12 mW	0.34 W	12–48 W
Frequency	5 MHz	10–11 MHz	20 kHz	100 kHz
Efficient	32–42%	25–48%	68%	75%

it is recent revolutions in computing and mobile devices that have stimulated the evolution of wireless power transmission technology. The three periods of wireless power transmission include the theory basement, large power transmission via microwave, and application on mobile devices. More mature solutions appeared from the late 1990s to the 2010s. The most sophisticated and widely used application is the Qi standard. Implantable medical devices can also be powered wirelessly. The unique design requirements inherent to medical applications present an opportunity to redesign the approach to best fit the medical field. In this chapter, we also proposed an approach to analyze the model of each parameter in the circuit and give detailed solutions to design a high-efficiency four coils transmission system. Also, in the last part of the chapter, we provided examples of implantable medical devices that take advantage of wireless power. Wireless power transmission may prove helpful in situations where the typical wired power approach is difficult to implement due to physical access limitations. Currently, there are barriers to be overcome in terms of efficiency and environmental inference. The increasing market share of commercial mobile devices and wireless power transmission itself will likely encourage more engineers to enter the field in the near future, and we hope that this influx of researchers in the area can contribute to the future of this technology.

ACKNOWLEDGMENTS

This chapter is made possible through the help and support from everyone, including my parents, teachers, family, friends, and, in essence, all sentient beings. The authors wish to specifically acknowledge Amber Bhargava, for her advice and assistance.

REFERENCES

[1] iPhone Battery with 1715 mAh Capacity Possibly Destined for ‘iPhone 6s’ or ‘6c’ Appears. [Online]. Available: http://www.macrumors.com/2015/08/03/iphone-6s-6c-1715mah-battery/. [Accessed: Sep 20, 2015].

[2] Apple, “Apple—iPhone 6—Technical Specifications”. 2015. [Online]. Available: https://www.apple.com/iphone-6/specs/. [Accessed: Sep 13, 2015].

[3] J. Pramis, Number of mobile phones to exceed world population by 2014, Digital Trends, 2013. [Online]. Available: http://www.digitaltrends.com/mobile/mobile-phone-world-population-2014/. [Accessed: Sep 13, 2015].

[4] J. Maxwell, A treatise on electricity and magnetism. New York: Dover Publications, 1954.

[5] H. Hertz, Electric waves. New York: Dover Publications, 1962.

[6] W. Brown, The history of wireless power transmission, Solar Energy, vol. 56, no. 1, pp. 3–21, 1996.

[7] W. Brown, Experiments in the transportation of energy by microwave beam, in 1958 IRE International Convention, New York, 1966, pp. 8–17.

[8] G. Goubau and F. Schwering, On the guided propagation of electromagnetic wave beams, IRE Transactions on Antennas and Propagation, vol. 9, no. 3, pp. 248–256, 1961.

[9] J. Degenford, M. Sirkis, and W. Steier, The reflecting beam waveguide, IEEE Transactions on Microwave Theory and Techniques, vol. 12, no. 4, pp. 445–453, 2015.

[10] G. Maryniak, Status of international experimentation in wireless power transmission, Solar Energy, vol. 56, no. 1, pp. 87–91, 1996.

[11] Koert P, Cha J T. Millimeter wave technology for space power beaming. IEEE Transactions on Microwave Theory and Techniques, 1992, 40(6): 1251–1258.

[12] Wang Zhixiong, Hu Jinlei, Liang Jun. Application prospect of wireless electric power transmission technology. Journal of Air Force Engineering University: Natural Science Edition, 2003, 4, 182–85 (in Chinese).

[13] Hirai J J, Kim T W, Kawamura A. Wireless transmission of power and information for cableless linear motor drive. IEEE Transactions on Power Electronics, 2000, 15(1): 21–27.

[14] Esser A, Skudelny H C. A new approach to power supplies for robots. IEEE Transactions on Industry Applications, 1991, 27(5):871–875.

[15] Manolatou C, Khan M J, Fan Shanhui, et al. Coupling of modes analysis of resonant channel add-drop filters. IEEE Journal of Quantum Electronics, 1999, 35(9): 1322–1331.

[16] Fredy S Q, Jesús G C, Jordi S, et al. Wireless power of single-chip systems with integrated coil and external wire-loop resonator. Applied Physics Letters, 2008, 92(7): 074102 (1–3).

[17] Tsang M, Psaltis D. Theory of resonantly enhanced near-field imaging. Optical Society of America, 2007, 15(19): 11959–11970.

[18] Kurs, A. Wireless power transfer via strongly coupled magnetic resonance. Science 2007, 317: 6–10.

[19] Ludwing R, Bretchko P. RF circuit design: theory and application[M]. Beijing: Science Press, 2002: 22–23.

[20] K. Vahalia, Wireless power charging and transmission market to Exceed $11B by 2020, Iotevolutionworld.com, 2015. [Online]. Available: http://www.iotevolutionworld.com/topics/smart-grid/articles/121877-wireless-power-charging-transmission-market-exceed-11b-2020.htm. [Accessed Sep 13, 2015].

[21] C.B. Teck, I. Takehiro, K. Masaki et al. Basic study of improving efficiency of wireless power transfer via magnetic resonance coupling based on impedance matching. Industrial Electronics(ISIE),2010 IEEE International Symposium, Bari, July 4–7, 2010: 2011–2016.

[22] Y.G. Su, C.S. Tang, S.P. Wu. Research of LCL resonant inverter in wireless power transfer system. International Conference on Power System Technology, Chongqing, October 22–26, 2006: 125–134.

[23] C. B. Zhu, K. Liu, C.L. Yu. Simulation and experimental analysis on wireless energy transfer based on magnetic resonances. 2008 IEEE Vehicle Power and Propulsion Conference, VPPC 2008, Harbin, 35–38.

[24] J.T. Boys, G.A. Covic, A.W Green. Stability and control of inductively coupled power transfer system. IEEE Proceedings: Electric Power Applications, 2000, 147(1):37–43.

[25] C. Byungcho, N. Jaehyun, C Honnyong, et al. Design and implementation of low-profile contactless battery charger using planar printed circuit board windings as energy transfer device. IEEE Transaction on Industrial Electronics. 2004, 51(1):140–147.

[26] M. Uman, D. McLain and E. Krider, The electromagnetic radiation from a finite antenna, American Journal of Physics, vol. 43, no. 1, pp. 33–38, 1975.

[27] R. Merrill and E. Issaq, Multi-turn, multi-level IC inductor with crossovers, US 5610433 A, 1997.

[28] Y. X. Xu, J. T. Boys. Modeling and controller design of ICPT pick-ups. Proceedings of the Power System Technology 2002: 1602–1606.

[29] A. Kurs, A. Karalis, M. Kesler, A. Campanella, K. Hall, K. Kulikowski and M. Soljacic, Low AC resistance conductor designs, US 8410636 B2, 2013.

[30] E. W. Baker and R. Sarpeshkar, Feedback analysis and design of RF power links for low-power bionic systems. IEEE Transactions on Biomedical Circuits and Systems, vol.1 March 2007; pp.28–38.

[31] R. Chambers, The anomalous skin effect, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 215, no. 1123, pp. 481–497, 1952.

[32] RamRakhyani, A.K.; Mirabbasi, S.; Chiao, M. Design and optimization of resonance-based efficient wireless power delivery system for biomedical implants. IEEE Transactions on Biomedical Circuits and Systems 2011, 5,48–63.

[33] Prnewswire.com, Global Qi Standard Powers Up Wireless Charging, 2015. [Online]. Available: http://www.prnewswire.com/news-releases/global-qi-standard-powers-up-wireless-charging-102043348.html. [Accessed: Aug 7, 2015].

[34] Wirelesspowerconsortium.com, Member list, 2015. [Online]. Available: http://www.wirelesspowerconsortium.com/member-list/. [Accessed: Aug 7, 2015].

[35] L. Mearian, Qi wireless charging spec gets more flexibility with distance extension, Computerworld, 2015. [Online]. Available: http://www.computerworld.com/article/2490634/emerging-technology/qi-wireless-charging-spec-gets-more-flexibility-with-distance-extension.html. [Accessed: Aug 5, 2015].

[36] T. Rob, Qi vs A4WP: War of the wireless charging standards, Android Authority, 2013. [Online]. Available: http://www.androidauthority.com/qi-a4wp-wireless-charging-standards-190836/. [Accessed: Aug 4, 2015].

[37] ZENS Wireless Charging, How does wireless charging work? | Qi Induction Technology, 2015. [Online]. Available: http://www.makezens.com/wireless-charging-how-does-it-work/. [Accessed: Aug 6, 2015].

[38] T. Rob, Wireless charging technology in 2015: what you need to know, Android Authority, 2015. [Online]. Available: http://www.androidauthority.com/wireless-charging-technology-2015-580015/. [Accessed: Aug 4, 2015].

[39] A. Maxham, Featured: top 10 best Qi enabled wireless chargers | Androidheadlines.com, AndroidHeadlines.com |, 2014. [Online]. Available: http://www.androidheadlines.com/2014/10/featured-top-10-best-qi-enabled-wireless-chargers.html. [Accessed: Aug 5, 2015].

[40] P. Sayer, Wireless power consortium to unleash electronic gadgets, PCWorld, 2015. [Online]. Available: http://www.pcworld.com/article/155766/article.html. [Accessed: Aug 6, 2015].

[41] Merger.powermatters.org, powermatters.org—Members: All Listings, 2015. [Online]. Available: http://www.merger.powermatters.org/index.php/component/mtree/members/all?sort=link_name&cf29=&cf33=&cf30=&link_name=&Itemid=120. [Accessed: Aug 8, 2015].

[42] Z. Honig, Duracell Powermat WiCC could bring wireless charging to any smartphone, we go hands-on (video), Engadget, 2015. [Online]. Available: http://www.engadget.com/2012/02/26/duracell-powermat-wicc-wireless-charging-card/. [Accessed: Aug 5, 2015].

[43] M. Joehren, K. Brink, K. Neumann, R. Dumont and M. Bruno, Development of an optimized wireless charging application solution, NXP Semiconductors, 2014. Available: http://www.nxp.com/documents/other/75017507.pdf. [Accessed: Feb 20, 2016].

[44] Maxwell, R., Wireless charging standards, Qi, Powermat, A4WP, what does it all mean and who will prevail?, *Phone Arena*, 2014. [Online]. Available: http://www.phonearena.com/news/Wireless-charging-standards-Qi-Powermat-A4WP-what-does-it-all-mean-and-who-will-prevail_id57371. [Accessed: Aug 7, 2015].

[45] E. Cervantes, A4WP and PMA merging in July—wireless charging is about to get much better, *Android Authority*, 2015. [Online]. Available: http://www.androidauthority.com/a4wp-pma-merging-wireless-charging-619240/. [Accessed: Aug 3, 2015].

[46] Szczesny S., Jetzki S., Leonhardt S., Review of current actuator suitability for use in medical implants, 28th Annual International Conference of the IEEE Digital Object Identifier, New York, NY, 2006: 5956–5959.

[47] Fda.gov, Syncardia Temporary CardioWest Total Artificial Heart (TAH-t)—P030011, 2013. [Online]. Available: http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/Recently-ApprovedDevices/ucm080816.htm. [Accessed: Sep 13, 2015].

[48] V.S. Mallela, V Ilankumaran, N. S. Rao, Trends in cardiac pacemaker batteries, Indian Pacing and Electrophysiology Journal, 2004 1(4):201–212.

[49] Guozheng Yan, Peng Zan, Dongdong Ye, et al., Design of transcutaneous energy transmission system for artificial anal sphincter, Proceedings of the 2007 IEEE International Conference on Mechatronics and Automation, China, 2007: 1434–1438.

[50] Lim H.G., Yoon Y.H., Lee C.W., et al., Implementation of a transcutaneous charger for fully implantable middle ear hearing device. Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 2005: 6813–6816.

[51] N. Chaimanonart, K.R. Olszens, M.D. Zimmerman et al., Implantable RF power converter for small animal in vivo biological monitoring, Proceedings of the IEEE Conference of Engineering in Medicine and Biology, Shanghai, China, 2005: 5194–5197.

[52] T. Takura, H. Ishiai, F. Sato, et al, Basic evaluation of signal transmission coil in transcutaneous magnetic telemetry system for artificial hearts, IEEE Transactions on Magnetics, 2005, 41(10), 4173–4175.

[53] K. Iwawaki, M. Watada, S. Takatani, et al., The design of core-type transcutaneous energy transmission systems for artificial heart. Proceedings of the 30th Annual

Conference of the IEEE Industrial Electronics Society, Busan, South Korea, 2004: 948–952.

[54] W. Ying, Y. Luguangm, and X. Shangang, Modeling and performance analysis of the new contactless power supply system. Proceedings of the 8th International Conference on ICEMS, Nanjing, 2005: 1983–1987.

[55] S. Arai, H. Miura, F. Sato, et al., Examination of circuit parameters for stable high efficiency TETS for artificial hearts, IEEE Transactions on Magnetic, 2005, 40(10): 4170–4172.

[56] S. Suzuki, T. Katane, H. Saotome et al., A proposal of electric power generating system for implanted medical devices, IEEE Transactions on Magnetics, 1999, 35(5): 3586–3589.

[57] J. Weber, K. Potje-Kamloth, F. Haase, et al., Coin-size coiled-up polymer foil thermoelectric power generator for wearable electronics, Sensors and Actuators A: Physics, 2006 132:325–330.

[58] S. Suzuki, T. Katane, H. Saotome et al., Electric power-generating system using magnetic coupling for deeply implanted medical, IEEE Transactions on Magnetics, 2002, 38(5): 3006–3008.

[59] J. Kymissis, C. Kendall, J. Paradiso et al., Parasitic power harvesting in shoes. Proceeding of Second IEEE International Conference on Wearable Computing, Pittsburgh, PA, USA, 1998: 132–139.

[60] C. B. Williams, R. B. Yates, Analysis of a micro-electric generator for microsystems, The 8th International Conference on Solid-State Sensors and Actuators, Sweden, 1995: 369–372.

[61] P. J. Glynne, S. P. Bebby, N. M. White, Towards a piezoelectric vibration-powered micro generators, IEEE Proceedings on Science Measurement and Technology, 2001: 68–72.

[62] K. P. Naresh, C. K. Vishrut, A high efficiency optical power transmitting system to a rechargeable lithium battery for all implantable biomedical devices, 3rd Kuala Lumpur International Conference on Biomedical Engineering, Kuala Lumpur, Malaysia, December 11–14, 2007: 533–537.

[63] G. Rosenberg, A. J. Snyder, W. J. Weiss, et al., A permanent left ventricular assist device: in vivo testing, IEEE Engineering in Medicine & Biology Society 10th Annual International Conference, New Orleans, LA, 1998: 65–67.

[64] K. Shiba, M. Nukaya, T. Tsuji, et al., Analysis of current density and specific absorption rate in biological tissue surrounding transcutaneous transformer for an artificial heart, IEEE Transactions on Biomedical Engineering, 2008, 55(1): 205–213.

[65] G. Wang, W. Liu, M. Sivaprakasam, et al., Design and analysis of an adaptive transcutaneous power telemetry for biomedical implants, IEEE Transactions on Circuits and Systems I, Regular Papers, 2005, 52(10):2109–2117.

[66] Watada M, Iwawaki K, Tamada T, et al. The development of core-type transcutaneous energy transmission system for artificial heart[C]. Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE, Shanghai, 2006: 3849–3852.

8

ENERGY-EFFICIENT PHYSICAL ACTIVITY DETECTION IN WIRELESS BODY AREA NETWORKS

Daphney-Stavroula Zois[1], Sangwon Lee[2], Murali Annavaram[3], and Urbashi Mitra[3]

[1]*Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
[2]*Computer Science Department, University of Southern California, Los Angeles, CA, USA*
[3]*Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA*

8.1 INTRODUCTION

Human health monitoring via eHealth—a term adopted for the first time back in 1999 [1] to encompass the range of services and systems at the intersection of medicine and technology—has been the center of attention for both research and commercial organizations. Google has recently announced the design of smart contact lenses for tracking glucose levels in diabetics, and Calico is a new company focusing on aging and related diseases. IBM recognizes health care as one of the five innovations that will change our lives within 5 years and focuses on DNA sequencing for cancer treatment. Qualcomm, Asthmapolis, and Zephyr are currently researching on asthma attack detection in children and teenagers using wearable technology. USC, Stanford, MIT, and Berkeley focus on a wide variety of techniques associated with health-related research, ranging from magnetic resonance imaging (MRI) techniques, new cancer treatment, and heart monitoring technologies, to techniques for helping the blind and automatic monitoring of induced comas. A driving force behind

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

these research initiatives is the increasing sophistication of wearable and implementable medical devices along with their integration with wireless technology, which has led to an ever-expanding range of therapeutic and diagnostic applications. Applications include, but are not limited to, physical activity detection for obesity prevention [2], patient rehabilitation [3], diagnostics, social communication and interaction development in autism [4], fall detection in elderly [5], and early detection, treatment, and prevention of diseases [6].

In many healthcare applications, accurate detection of a person's state is critical and can be used to measure an individual's physical activity level. Properly assessing physical activity levels in individuals can prevent various conditions and diseases, such as obesity, type 2 diabetes, cardiovascular diseases, and various types of cancers, as well as promote physical and mental health [7]. In this chapter, our focus is on the development of systems that continuously monitor and interpret physical activity with minimum intervention to address obesity, which has taken the form of an epidemic in our society [8, 9]. Previously, various approaches have been considered to enable continuous detection and recording of physical activity, since the latter constitutes a particularly difficult problem [7]. However, there is no single accepted methodology for physical activity detection, while individual-specific factors such as gender, amount of body fat versus lean tissue, and age significantly affect detection, further complicating the situation. It is common in preventive health applications to employ accelerometer (ACC) systems. Unfortunately, these system face numerous challenges, for example, complicated calibration methodologies, data downloading, processing, and interpretation must be done by experts; comparison between proprietary systems and further signal processing cannot be accomplished [7]. In contrast to these, we have implemented the KNOWME system [10], a prototype wireless body area network (WBAN), which alleviates the problems discussed before by enabling real-time monitoring of physical activity and adaptive, personalized interventions.

WBAN technology is an emerging research area that uses a combination of health sensors and personal devices along with their wireless communication capabilities to enable continuous body monitoring. A typical WBAN, illustrated in Figure 8.1, is equipped with a limited number of light-weight, low-power, nonobtrusive heterogeneous sensors (e.g., ACC, electrocardiographs (ECGs), and pulse oximeter), which are attached to clothing or on the body, and a fusion center, which is usually an energy-constrained personal device, for example, mobile phone and personal digital assistant (PDA). WBANs have enabled a wealth of applications in health care, sports, entertainment, lifestyle, military, and emergency situations. Having the potential to revolutionize health care, WBANs need to be able to support data collection, efficient decision making and interaction with the individual, the cloud, and healthcare professionals toward improving each individual's lifestyle.

Due to their unique nature, practical realization of WBANs faces a lot of distinctive challenges, such as sensor sensitivity, unobtrusiveness, reliability, real-time operation, security, privacy, user friendliness, and standardization [11]. The development of the KNOWME system and the associated in-lab deployment studies revealed the following three major research challenges:

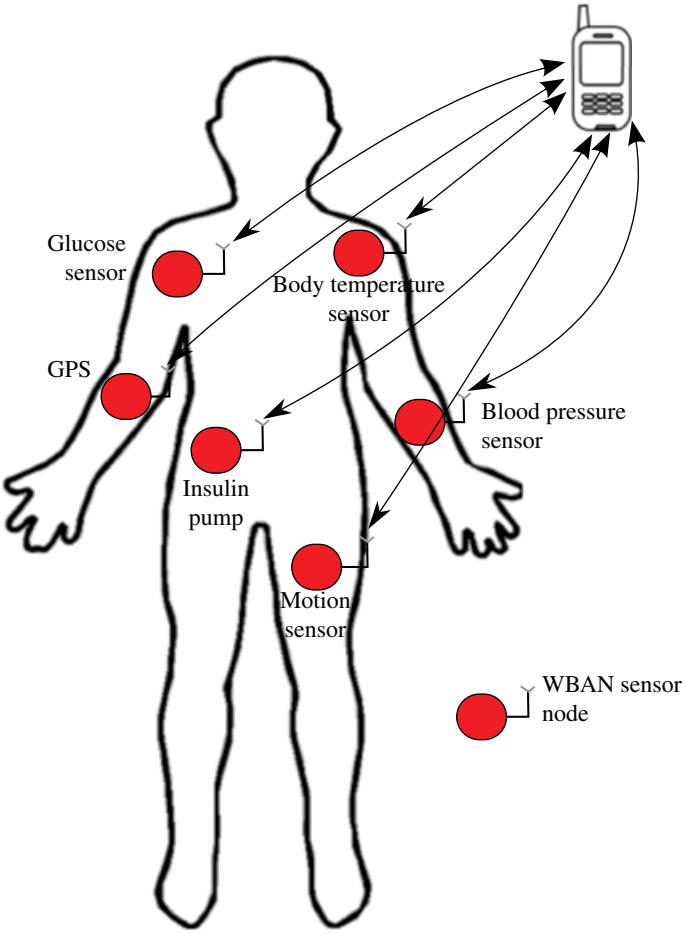


FIGURE 8.1 WBAN schematic.

1. *Energy efficiency.* Uninterrupted collection of high fidelity data requires the continuous communication between the fusion center and the biometric sensors, which can dramatically decrease its battery life. Thus, it is imperative to devise efficient algorithms that balance high-accuracy physical activity detection and the need for energy efficiency.
2. *Robust system design.* The system architecture needs to be carefully designed to accommodate potential application instabilities and slowdowns due to complex signal processing operations as well as mobility of individuals and incorrect sensor placement.
3. *Accurate physical activity detection and classification.* To account for multiple heterogeneous data, inconsistent sensor placement, and intersession variability, it is necessary to devise novel signal representations and classification algorithms and employ personalized training.

The focus of this chapter is the energy-efficient operation of the KNOWME network, which constitutes a significant factor in the long-term deployment of WBANs. As WBANs operate continuously, one of the major bottlenecks to data collection and analysis is the limited battery life of mobile phones. For instance, the authors in Ref. [12] report that a Nokia N95 phone battery has nearly 200 h of standby time, but when it is used for user state detection, the battery drains in 4 h. This chapter tackles the energy consumption challenges in WBANs using a two-step process.

For WBANs to be used pervasively, it is necessary to first understand where the energy is spent in a WBAN and then use energy-efficient strategies to operate it. As such in the first part of this chapter, we present a detailed energy usage characterization of a WBAN. At each stage of a WBAN design and operation, the system designer has a wide range of options to trade off battery life for other important metrics. For example, during system software development phase, the designer has a choice of programming platforms for improving software productivity. The designer may use Python, Java, or a phone's native programming model such as Symbian C++ or iPhone SDK to develop the system software. The choice of language can trade off the software productivity with potential runtime overheads of managed environments that lead to energy inefficiency. During the data transmission phase, the designer has to make energy trade-off decisions on computation and communication costs. Storing data on flash may allow the mobile phone to compress large chunks of data and send compressed data to the remote server. Data compression places more computation demands. On the other hand, compression may reduce the transmission energy cost. Local storage of data also allows the mobile phone to process the data locally and only send interesting events to the back-end. For instance, detecting an abnormal heart beat signal from ECG will reduce the need to continuously transmit normal operational data thereby reducing transmission energy. As demonstrated with these ample examples, the system designer is faced with a daunting list of trade-offs. While some previous studies have quantified energy trade-offs in specific domains, such as communication versus computation [13, 14], sampling rate versus accuracy [15–17], to the best of our knowledge, there is no published work that provides a comprehensive quantification of energy costs of the various choices a designer has to make.

In the second part of this chapter, we tackle the challenge of optimizing energy consumption due to sensor data collection. In particular, we study the problem of optimizing the resource allocation strategy among WBAN heterogeneous sensors employed by the fusion center to enable energy-efficient activity detection. As we will see later on, the adopted strategy significantly affects both detection accuracy and energy consumption since each sensor's discriminative capabilities vary per activity, while incurring different energy costs. The interperson variability also affects the selection of the appropriate strategy, suggesting the need for personalization.

The remainder of this chapter is organized as follows. Section 8.2 describes the architecture of the KNOWME network, while Section 8.3 discusses the energy impact of various design choices. Next, Section 8.4 formalizes the problem of physical activity detection in energy-constrained WBANs, and Section 8.5 discusses optimal and suboptimal sensor selection strategies. In Section 8.6, a more advanced

formulation of the sensor selection problem is presented, and in Section 8.7, optimal and suboptimal algorithms are proposed. Section 8.8 illustrates the performance of the sensor selection strategies on the experimental data collected via the KNOWME network. Section 8.9 summarizes prior work on energy-efficient algorithms in sensor networks, hand-held devices, and WBANs. Finally, Section 8.10 concludes the chapter.

Throughout this chapter, we adopt the following notational conventions. Unless stated otherwise, all vectors are column vectors denoted by lowercase boldface symbols (e.g., $\mathbf{v}$), and all matrices are denoted by uppercase boldface symbols (e.g., $\mathbf{A}$). Sets are denoted by calligraphic symbols (e.g., $\mathcal{X}$). $\mathbf{1}$ denotes a vector with all components equal to 1, $\mathbf{0}$ is a vector with all components equal to 0, and $\mathbf{I}$ is the identity matrix with dimensions determined from the context. tr$(\cdot)$ denotes the trace operator and diag$(\mathbf{x})$ the diagonal matrix with elements of the components of vector $\mathbf{x}$. Finally, for any event B, $\mathbf{1}_B$ denotes the corresponding indicator function, that is,

$$\mathbf{1}_B = \begin{cases} 1 & \text{if } B \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

8.2 KNOWME PLATFORM

In this section, we provide details of the KNOWME WBAN platform. KNOWME is a low-cost mobile phone-centric WBAN, which is currently deployed in Los Angeles for pediatric obesity prevention and treatment. In the current implementation, each KNOWME node consists of a Nokia N95 mobile phone (fusion center) and a set of sensors: triaxial ACCs that measure motion, an ECG sensor that records heart rate, a blood oxygen saturation sensor (OXI) that monitors blood oxygen levels, and the Global Positioning System (GPS). The external ACC, ECG, and OXI are Bluetooth-enabled off-the-shelf sensors from Alive Technologies, while N95 includes a built-in ACC and a GPS. The ECG samples at 300 Hz, OXI uses 100 Hz sampling, the external ACC samples at 75 Hz, the internal ACC uses 30 Hz sampling, while GPS can be sampled at any user-specified rate.

The KNOWME software runs as a mobile application on N95 that has two components: (i) a background process (KMCore) for data collection and (ii) a client interface application (KMClient) for sensor configuration and data visualization. The KMCore comprises seven components arranged in a four-tier hierarchy: (i) device manager at the bottom-most tier, (ii) data collector at the second tier, (iii) data analyzer, local storage manager, and data transmitter at the third tier, and (iv) a service manager at the top tier. Figure 8.2a shows how various components in the KMCore interact with each other. There is one thread per each sensor referred to as device manager thread that receives data from its associated sensor. By creating a separate thread per each sensor, KNOWME continues to receive data from working sensors without being hindered by a single nonfunctional sensor. The data collector thread receives sensor data from each device manager and synchronizes all the sensor data with the same timestamp into a single health record. It combines several

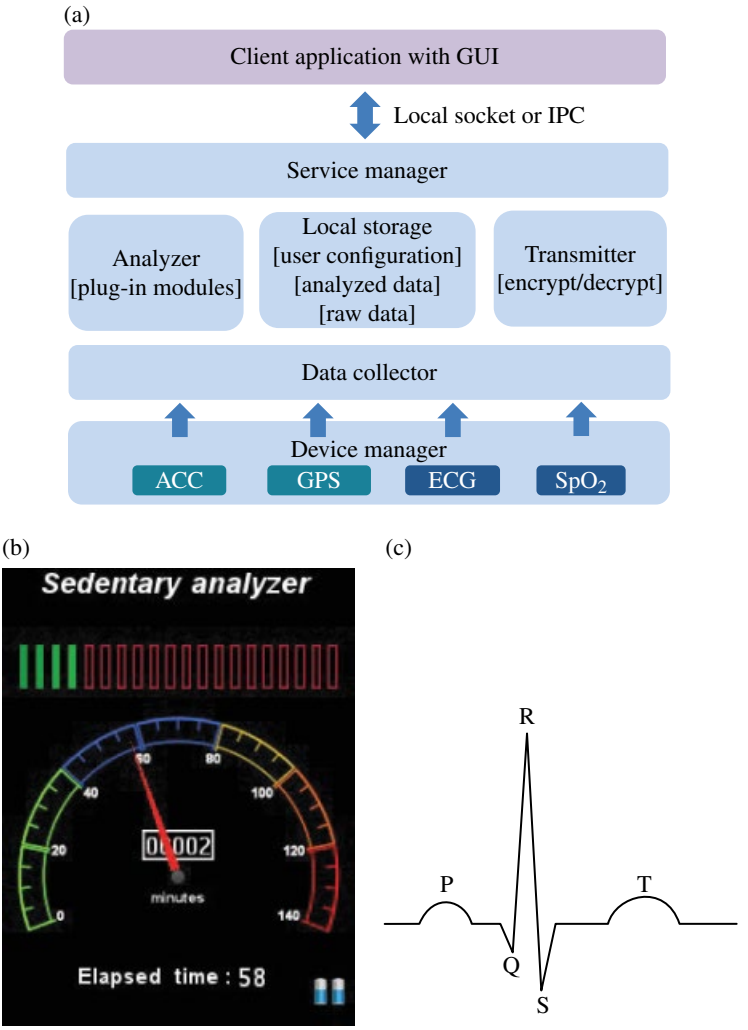


FIGURE 8.2 KNOWME application: (a) the components of KMCore, (b) a screenshot of KMClient running on N95, and (c) ECG signal.

such records into a single write buffer and sends it to the local storage manager to write the data to the flash storage. When the flash storage runs out of space, old data is replaced with new data in a first-in, first-out manner.

The analyzer modules are designer-defined modules that perform domain-specific tasks. In pediatric obesity, these modules perform activity classification using multimodal signal-processing algorithms. In the current implementation, the analyzer classifies an activity as either sedentary or nonsedentary using just ACC data along with a sedentary analyzer (SA) based on a support vector machine classifier [18]. The transmitter module transfers data to the back-end and handles data compression and encryption for privacy and energy saving.

The back-end server runs a more comprehensive suite of classification algorithms to detect a range of user activities such as walking, running, fidgeting, standing, and sitting using ECG and ACC sensor data with multimodal signal processing [19]. Such finer classification is used by physicians at the back-end to get a comprehensive understanding of user behavior and to precisely measure the calories burned. The back-end server stores the sensor data indefinitely. In other words, the data stored on the back-end server is a complete record of a person's physical activity history, and the most recent time window of this history is stored on the mobile phone's flash storage.

The last component of the KMCore application is a service manager thread that uses sockets or interprocess communication (IPC) to provide the sensor data to other mobile applications running on the phone such as the data visualization application. Figure 8.2b shows a visualization screen of the KMClient that shows how long a user is sedentary.

8.3 ENERGY IMPACT OF DESIGN CHOICES

Based on the framework described in the previous section, we will now provide a comprehensive evaluation of the energy consumption of the various KNOWME components. While KNOWME is one particular implementation of a WBAN, we would like to note that most WBANs that we are aware of have very similar architecture [20, 21]. Furthermore, wherever possible, we quantify the energy consumption of basic operations without relying on KNOWME semantics. For instance, energy consumed for a byte of data compression and transmission is independent of whether it is performed within the KNOWME framework or otherwise. In each section later, we evaluate the energy cost of each KNOWME design component exploiting multiple available choices.

8.3.1 Impact of Software Platform on Energy Consumption

Selecting an appropriate software development platform is arguably the most important factor in any system design. The programming platform choice can determine the development cost in terms of person-hours as well as the system performance. Hence, in this section, we first evaluate the design choices available for mobile phone software development in terms of their energy efficiency. There are three popular SDKs for programming Symbian Operating System-based phones such as N95: Symbian C++, Java 2 Micro Edition (J2ME), and S60 Python (PyS60), and we illustrate their relative energy efficiency. For Android and iPhones, the programming interfaces are similar to the J2ME interfaces on Symbian.

We have selected three functions that are of particular interest to WBANs: a heartbeat detection algorithm (QRS detection [22]), advanced encryption standard (AES) encryption, and GNU zip (Gzip) data compression. AES encryption is commonly used to transmit data from the mobile phone to the back-end server for data protection. Gzip is also another commonly used function to compress data before transmitting them to the back-end server to reduce transmission costs. As shown in Figure 8.2c,

the ECG signal is characterized by the P, QRS, and T waves. The QRS detection algorithm (QDA) detects R peaks that are used as a basis of reference in ECG segmentation and are necessary to recognize heartbeat. We have used the QDA implementation available from the Open Source ECG Toolbox [23] and have ported it to each of the three programming languages.

The first set of nine bars in Figure 8.3 illustrate the execution time of the three functions using the three programming platforms (the remaining bars will be discussed in Section 8.3.2). In Table 8.1, the row with label N95 shows the corresponding energy consumption in Joules reported by the Nokia Energy Profiler Tool [24], which is fairly lightweight and adds negligible overhead. We have used 10 min of ECG data that correspond to 180 kB of data, as input to each of the following three functions: QDA, AES, and Gzip. As Figure 8.3 illustrates, PyS60 performance on AES is three orders of magnitude worse than C++ implementation, since PyS60 does not have a just-in-time (JIT) compiler unlike PCs. However, for QDA, PyS60 is about 40 times slower than Symbian C++ (comparing Bars 1 and 3 in Fig. 8.3) and has a corresponding 40 times higher energy consumption (see Columns 1 and 3 in Table 8.1). The improved performance of PyS60 in QDA case is attributed to the efficient implementation of several built-in library functions, such as low-pass filtering in PyS60. For Gzip, PyS60 performed as fast as Symbian C++. Further analysis of PyS60 Gzip function showed that PyS60 implements Gzip as a native function written in Symbian C++ and included in PyS60 as an extension module. In essence, PyS60 implementation of Gzip is a Symbian C++ implementation of Gzip. Note that qualitatively these results should come as no surprise; interpretive languages are slower than native execution. But quantitatively, differences between various programming models are quite significant, and a mobile applications programmer must be made aware of these vast differences.

8.3.2 Sensitivity to Hardware Platform

One may raise the obvious concern if the results in the previous section are specific to the underlying hardware platform of N95. *What if these applications behave differently on a different mobile phone?* In order to clarify these concerns, we repeated the same set of experiments on a Nokia E75 and an Apple iPhone. Table 8.2 provides the relevant hardware and software specifications where the mobile phones differ. All platforms are based on ARM cores, but the iPhone processor is nearly twice as fast as the other two and has also twice the amount of memory. Hence, memory-related issues like garbage collection are likely to be more severe on Nokia platforms compared to iPhone. Furthermore, the iPhone provides only one programming interface, namely iPhone SDK, which is closer to the Symbian C++ in terms of programming complexity.

The second set of nine bars in Figure 8.3 illustrate the execution time of the three functions while running on E75. The last set of three bars in the same figure show the execution time of the three functions while running on iPhone. Similarly in Table 8.1, rows labeled E75 and iPhone show the corresponding energy consumption in Joules. Comparing results between N95 and E75, we observe that the relative impact of programming language on execution time (and energy) across all applications remains the same. In addition, execution time decreases in almost all cases on E75.

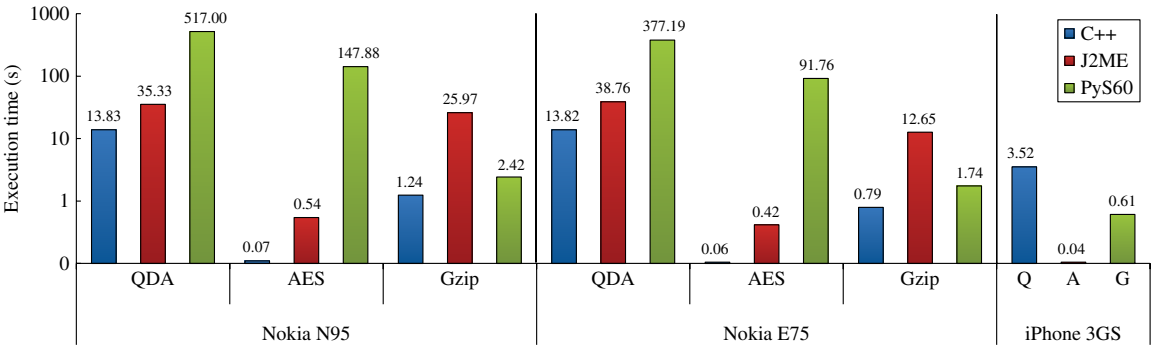


FIGURE 8.3 Execution time of three WBAN functions on three programming languages.

TABLE 8.1 Energy Consumption for Three Processing Functions

	QDA (J)			AES (J)			Gzip (J)		
Model	C++	J2ME	PyS60	C++	J2ME	PyS60	C++	J2ME	PyS60
N95	6.6	19.34	270.97	0.03	0.30	77.62	0.50	12.22	1.35
E75	5.75	15.95	151.63	0.02	0.16	36.89	0.33	4.93	0.70
iPhone 3GS	2.58	n/a	n/a	0.04	n/a	n/a	0.55	n/a	n/a

n/a, not applicable.

TABLE 8.2 Specification of the Mobile Phones

Spec	N95	E75	iPhone 3GS
CPU	Dual ARM 11 332 MHz	ARM 11 369 MHz	ARM A8 600 MHz
Memory	128 MB RAM	85 MB RAM	256 MB RAM
OS	Symbian S60	Symbian S60	iOS4

The primary reason for the improved performance and energy efficiency on E75 is that these applications are all single-threaded and the higher frequency ARM 11 processor on E75 executes them faster. The iPhone executes these applications even faster compared to Symbian C++ implementation on E75. Again, the reason is that the iPhone processor is nearly two times faster than E75. All these applications have similar code footprint, given that all platforms use ARM ISA. Hence, the relative execution time differences of the three applications remain the same.

8.3.3 Energy Consumption of Sensors

After establishing the energy impact of software development choice, we now focus on the energy costs of sensing itself. In the KNOWME network, built-in sensors such as ACC consume energy to perform the sensing operation, while external sensors (ECG and OXI) cause battery drain in mobile phones due to Bluetooth communication. Table 8.3 shows the energy consumption of the built-in phone sensors and the energy consumption when reading data from external sensors using Bluetooth. This data is obtained while sensing for a 10 min interval using the sampling rates shown in Table 8.3. A Symbian C++ application is used for reading the sensor information. ECG generates 300 samples per second, while OXI generates 100 samples per second. However, these sensor samples are internally buffered in the sensor and transmitted in bulk over the Bluetooth channel. Although ECG takes 300 samples, it only transmits four packets per second with each packet having 75 sensor samples. OXI transmits 10 packets per second. Hence, the energy consumption while receiving data from the OXI sensor is slightly higher than when communicating with ECG. Interestingly, when both ECG and OXI concurrently send data, the mobile phone is more energy efficient since the energy expenditure is much less than the sum of the energy spent when both sensors transmit data in isolation. The energy consumed in putting the Bluetooth radio in an active listening mode is amortized

TABLE 8.3 Energy Cost of Sensor Readings

Sensor	Energy (J)	Sampling Rate	Transmission Rate
Built-in ACC	37.804	30	30
ECG	114.846	300	4
OXI	137.433	100	10
ECG and OXI	156.419	300 and 100	4 and 10
Assisted GPS	53.994	0.003	0.003

over both sensors when hearing from the two sensors simultaneously. The built-in ACC consumes 38 J for 10 min while generating 30 samples per second.

8.3.4 Energy Consumption of GPS

Table 8.3 also shows the GPS energy consumption of the assisted GPS technology used in N95. Two GPS readings are made in the 10 min interval, and hence the sampling rate is 0.0033 samples per second. As seen, GPS consumes significantly higher power per each sample than any of the other sensing functions in KNOWME. Given that GPS is an energy-intensive operation, mobile phones provide various options for obtaining location information. In particular, there are four choices for the N95: (i) Bluetooth-based external GPS, (ii) assisted GPS, (iii) traditional GPS with no assistance, and (iv) network-based GPS that uses only cell towers to provide approximate position information. Figure 8.4 illustrates the power consumption trade-offs and the time to obtain the first GPS reading for each of these four methods. Generally, once the position is known, the next GPS reading cost is reduced. For example, in the assisted GPS mode, the first reading cost is 29 J per sample, whereas the second reading reduces to 25 J. In Figure 8.4, the Bluetooth-based external GPS power consumption shows only the mobile phone power consumption for establishing a Bluetooth link and reading the GPS coordinates. We observe a spike in the power consumption when the connection is established and the data is read over the Bluetooth channel. The power consumption drops down to the idle power when Bluetooth is ON but not actively sending/receiving data. The assisted GPS curve shows a large power spike when it communicates with the cell tower to get the rough location first. Once the cell tower provides the orbital data of GPS satellites, the GPS receiver on the mobile phone can narrow the search for satellite signals and quickly obtain the position information. In our measurements, this approach required roughly 15 s to receive the position information from the network and another 35 s to retrieve the precise position from a cold start. The integrated GPS curve corresponds to the basic nonassisted GPS mode. In this mode, the GPS receiver continuously scans for the satellite information. This approach uses lower peak power but takes nearly 100 s before obtaining the position information due to the absence of any assistance from the cellular network; the total energy consumed by integrated GPS is 33.9 J. The last curve shows the power consumption of network-based position information which, like assisted GPS, communicates with cell tower to triangulate its approximate position information, but no further data is retrieved from satellites. Hence, the

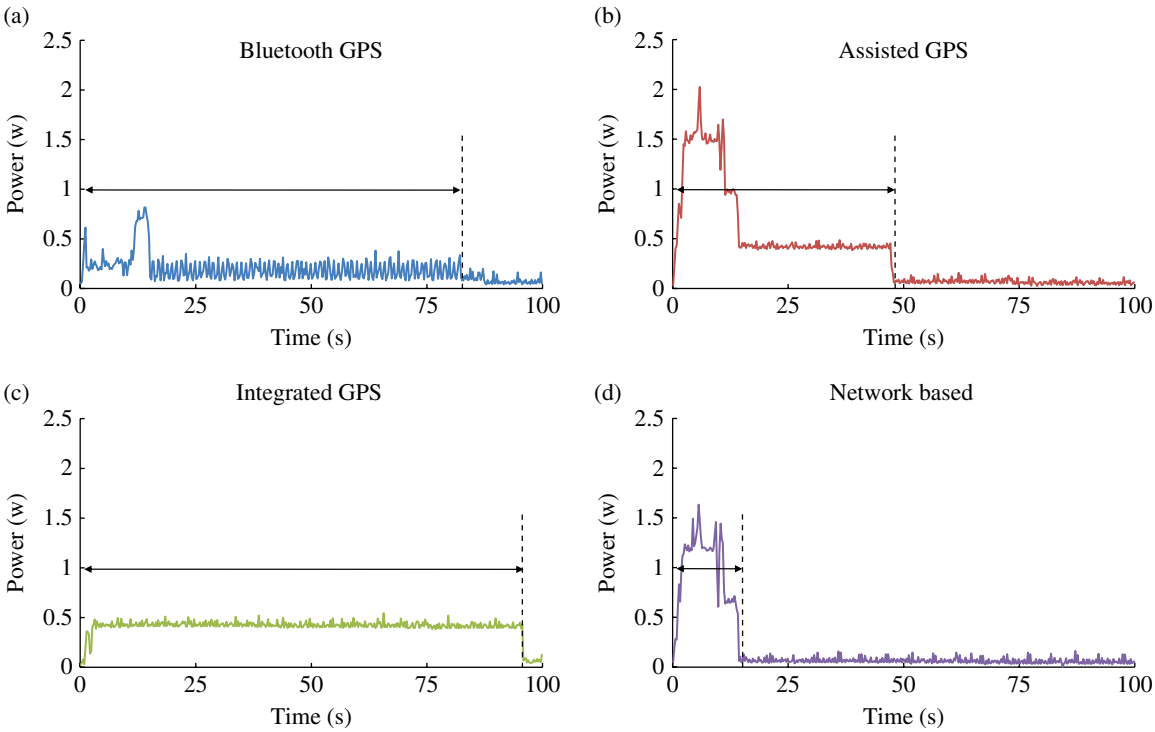


FIGURE 8.4 Energy cost of positioning methods.

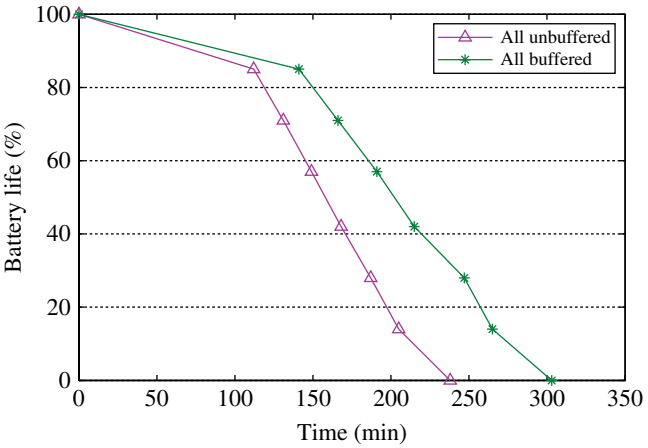


FIGURE 8.5 Battery drain with storage.

network-based GPS mode also consumes roughly the same power as the assisted GPS initially when communicating with the cell tower for 15 s. The consumption then drops to idle power since it does not try to compute accurate position information by communicating with satellites. Based on the total energy consumption (i.e., the product of the power consumption and the time to get a GPS reading), our evaluations show that the network-based GPS corresponds to the best choice for saving energy in a WBAN (13 J per sample) even though location data is only approximate. The network-based GPS is more than two times energy-efficient than the assisted GPS. When precise location is needed, then the assisted GPS is the best option.

8.3.5 Storage Costs

Once the sensor data is received, the mobile phone may write the data to the phone's flash memory for further analysis. Figure 8.5 shows the battery level of the mobile phone as we continuously sense data from ECG and OXI and write sensor data to the flash memory. The ALL_UnBuffered curve shows the battery level on the mobile phone as we write each packet of data immediately to the local flash without any buffering. In other words, every sensor sample received is immediately written to the flash memory. It is well known that flash energy efficiency is significantly compromised for small writes. Flash writes must be done at the size of a page granularity, typically 4 kB pages. If a smaller than page size write is performed, it is usually the case that the page that is being modified must be first read from the flash into a DRAM buffer and the bytes that are going to be written are updated in the DRAM buffer. Then, the entire DRAM buffer is written back to flash. Hence, writes lead to a read–modify–write sequence in the Flash memory, where even a few bytes of write translate to a full page write, which is referred to as write amplification effect [25].

The ALL_Buffered curve shows the battery level if we buffer the writes to DRAM and send large chunks to write to the flash. In this case, we buffer sensor data until

we receive at least 100 packets from a sensor. We then write the buffered data to the flash. As it can be seen, buffering improves the battery life from 240 to 299 min. Note, however, that the battery lifetime here does include the cost of using Bluetooth to receive data as well. Hence, the 240 min of battery life is not just due to flash writes only. Rather, we focus on the difference in the battery lifetime with and without buffered writing to the flash versus the absolute values. The difference of 50 min translates into an additional 25% increase in the battery lifetime in this experiment.

8.3.6 Computation versus Communication

As already discussed in Section 8.2, the N95 mobile phone does not only perform data collection. Instead, the analyzer module can perform local data analysis of the sensed data, and a transmitter module transmits data to the back-end server, optionally using compression and/or encryption. Therefore, the last trade-off we consider in this study is the energy consumption cost of performing local data analysis versus performing data analysis on the back-end server with the additional cost of data communication energy. This computation versus communication cost is not unique to KNOWME, as we expect most WBANs to make this fundamental trade-off.

8.3.6.1 Communication Costs To quantify the energy costs of communication, we have created a testbed. The testbed can send data from the mobile phone to the back-end server using either 3G, EDGE, or Wi-Fi. The AT&T broadband network is used for 3G and EDGE, while an 802.11g Linksys router is used for Wi-Fi. We have varied the data size transmitted to the back-end server from 100 to 1000 kB. As the performance of mobile wireless networks varies from place to place and from carrier to carrier, we have tested all data transfers from the same location during early morning within 1 h to reduce network congestion problems with other users. We have repeated this study multiple times always at the same time of the day to quantify day-to-day variations.

A data transmission consists of three phases: a connection phase, a transfer phase, and a tail phase. Figure 8.6 shows the power consumption during the three phases for the three wireless data transmission approaches. In this experiment, we used a 200 kB data transfer. Due to short range, even though Wi-Fi has much higher bandwidth, the peak power consumed is only slightly worse than 3G and EDGE radios on N95.

Figure 8.7 shows data transmission (uplink) energy costs of the three wireless interfaces as we increase the size of data transfer from 100 to 1000 kB. Wi-Fi is the most energy efficient across all data packet sizes. Hence, the overall energy consumption using Wi-Fi is significantly less than either 3G/EGDE in our setup. Obviously, due to limited Wi-Fi coverage, a practical WBAN implementation will most likely use either 3G or EDGE for real-time data transfer for mobile users. In this experiment, 3G consumes more energy than EDGE until about 400 kB of data size. In order to understand the reason, Table 8.4 shows average connection and tail energy costs for all three network interfaces. The connection and tail energy of 3G are much higher than EDGE, while the transfer energy is lower. Thus, for small data packets,

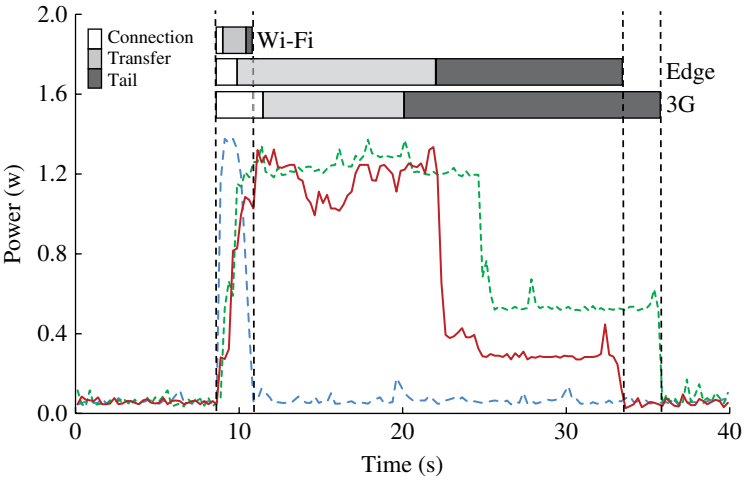


FIGURE 8.6 Three phases of transmission.

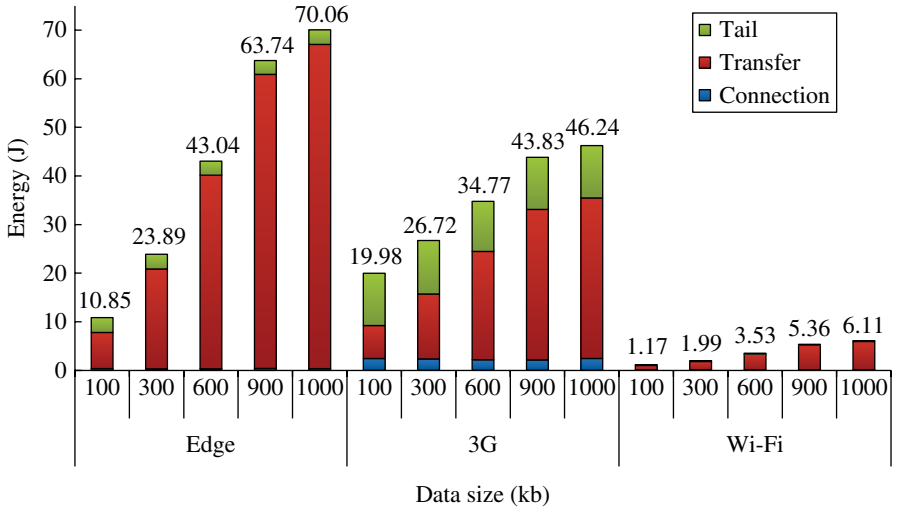


FIGURE 8.7 Data transmission costs (uplink).

TABLE 8.4 Connection and Tail Energy Costs

Medium	Connection Energy (J)	Tail Energy (J)
EDGE	0.346	2.987
G	2.331	10.752
Wi-Fi	0.132	0.166

3G consumes more energy than EDGE. As the data packet size increases beyond 400 kB, the transfer energy dominates the overall energy costs, and hence 3G consumes less energy than EDGE.

8.3.6.2 Local or Remote Computation In the KNOWME network, the most complex data analysis function is to detect user state to identify long phases of physical inactivity (sedentary behavior). The choice of whether to perform the user state detection on the mobile phone or on the back end depends on the total energy cost, and it is necessary to consider several factors. Consider a simple case, where we are interested in performing QDA on 10 min of ECG data, that is, 180 kB of data. Figure 8.8 shows the energy cost of this computation for local and remote computation. The first two bars in the graph show the energy cost of local computation when QDA is implemented in C++ or J2ME. The second set of three bars shows the transmission cost to perform data analysis on the remote server using three different transfer options: EDGE, 3G, and Wi-Fi. The last set of six bars shows the energy cost when data is compressed and then transmitted. The compression algorithm is implemented in C++ and J2ME and the transfer options are EDGE, 3G, and Wi-Fi.

Let us consider the J2ME implementation of QDA. The local computation cost is 19.34 J. The remote computation cost (without Gzip) varies between 1.46 J using Wi-Fi up to 22.72 J using 3G. On the other hand, the remote computation cost with Gzip implemented in J2ME varies from 13.46 J using Wi-Fi to 32.89 J using 3G. Now consider C++ implementation of the QDA. The local computation cost is 6.6 J. The remote computation cost without Gzip remains the same as before since there is no software platform dependence on transmission cost. But when Gzip is also implemented in C++, the energy cost varies between 1.74 J using Wi-Fi up to 21.17 J using 3G. When Wi-Fi is available, it is clearly energy efficient to perform remote computation. But when the user is roaming, which will be a common case in WBAN

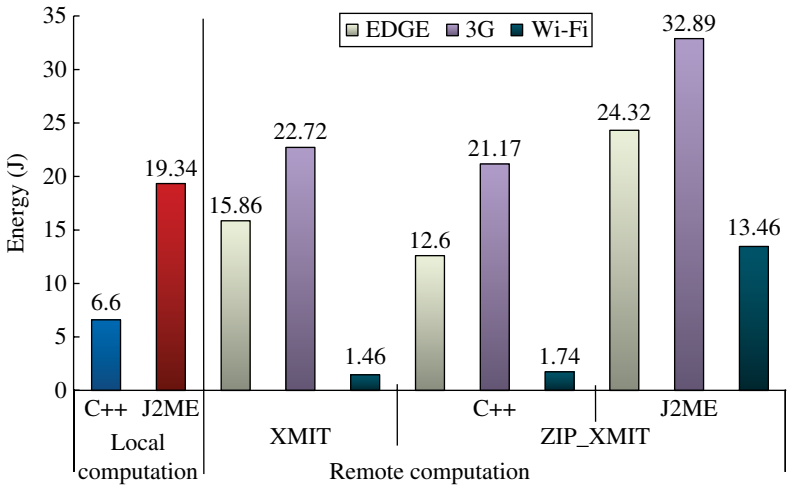


FIGURE 8.8 Local versus remote computation of QDA.

operation, local computation is better than EDGE or 3G cost when QDA is implemented in C++. But if QDA is implemented using J2ME, then the remote computation using EDGE and without Gzip compression (15.86 J) is better.

Even in this simple scenario, the choice of remote versus local computation is a complex function of software platform chosen, wireless radio used, and whether or not data is compressed. We even simplified the discussion by removing the network signal quality issues that may alter the energy costs dynamically; as explained earlier, in our setup, we have used the network during the least congested time and from a location with the best signal quality. Through this simple experiment, we demonstrate that there is no single statically best choice when it comes to trading off energy costs of computation with communication.

8.3.7 Summary

In this section, we summarize our results and conclusions from the WBAN characterization experiments. In particular, the following:

- It is qualitatively obvious that interpretive languages are likely to be slower than languages that run natively. However, there is a growing trend in traditional computing to use interpretive languages due to their programming simplicity and portability. Nonetheless, in the WBAN domain, the stringent battery constraints suggest that native execution provides significant energy savings.
- It is more energy efficient to communicate with multiple Bluetooth-enabled sensors concurrently rather than sequentially. Concurrent connections amortize the energy cost of the connection establishment cost.
- To enable localization, we considered four different types of services with different energy and location accuracy trade-offs. This trade-off is an application-dependent choice, but our results show at least two times improvement when using network-based GPS compared to assisted GPS.
- Buffering data in the flash storage before transmission adds a small delay before the data is received at the back-end server. Even for WBANs that real-time data is necessary, the additional delay due to buffering is minimal while it improves energy efficiency by at least 25%.
- As far as data transmission is concerned, Wi-Fi exhibits the lowest energy cost. On cellular network, however, it is unclear that 3G has always higher cost than 2G since for small data packets the 2G network is more energy efficient that 3G due to smaller connection and tail costs.

It is clear from the study that sensor management is key for improving the energy efficiency of WBANs. In the next section, we tackle one specific problem of sensor management, namely, optimizing sensor sampling rates to match specific user state detection needs. By potentially reducing sensor samples the amount of data that is collected, transmitted, and processed can be curtailed, which in turn can reduce the

energy consumption of the WBAN. We next describe the problem formulation, which is kept as generic as possible to encompass any WBAN, for optimizing the sensor sampling rate.

8.4 PROBLEM FORMULATION

We consider K heterogeneous, commercial sensors in a star topology and an energy-constrained mobile phone in a WBAN setting. An individual wearing the WBAN is alternating between a set of prespecified physical activites, for example, Sit, Stand, and Walk, and a set of biometric signals is generated by WBAN sensors and communicated to the mobile phone via Bluetooth. The goal is to determine the physical activity performed by the individual at each step.

As already discussed, the mobile phone undertakes coordination, sample collection, processing, and computation. In contrast, WBAN sensors simply transmit samples to the mobile phone using the Bluetooth protocol. As a result, the energy-constrained budget of the mobile phone, and not the sensors, limits the lifetime of the WBAN. This is a direct outcome of the reception of samples from the sensors being more energy costly than the transmission of such samples from the sensors. Undoubtedly, collecting equal number of samples from all WBAN sensors leads to undesirably high energy consumption, that is, a fully charged mobile phone battery is depleted in a couple of hours in contrast to WBAN sensors that can run for a couple of weeks. Fortunately, internal sensing energy costs are less than Bluetooth communication costs, and selecting to receive less than the available samples from each sensor in conjunction with the fact that certain sensor types are more informative in discriminating between certain physical activities can lead to significant energy savings. Nonetheless, if we restrict to internal sensors and/or few samples, there is a risk of substantial performance degradation. This inherent trade-off leads to a very interesting research problem: how to maximize the lifetime of such a unique sensor network, while ensuring acceptable detection performance. Our goal is to accurately track the unknown, time-evolving physical activity of the individual, while at the same time minimizing the reception cost. To this end, we want to construct efficient algorithms that optimize the listening schedule of the mobile phone: which sensor to listen to and for how long. Thus, we design a method that decides which WBAN sensors the mobile should listen to (if any, then one at a time) and the number of samples to receive from each sensor. We start with the stochastic model of our system.

8.4.1 Stochastic Model

We consider a dynamical system, where time is divided into discrete time slots and $k = 0,1,\ldots$ denotes discrete time. We define as system state the individual's current physical activity, which we denote at time step k by $\mathbf{x}_k$. The physical activity can take n possible values, that is, $\mathcal{X} \triangleq \{\mathbf{e}_1,\ldots,\mathbf{e}_n\}$ with $\mathbf{e}_i$ denoting the n-dimensional unit vector with 1 in the i-th position and zeros elsewhere, and each such vector indicates a different activity, for example, $\mathbf{e}_1$: Sit, $\mathbf{e}_2$: Stand, etc. We model the emporal

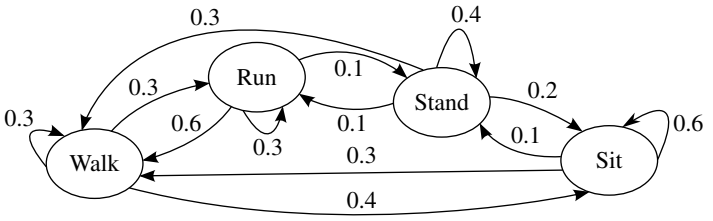


FIGURE 8.9 Markov chain of four physical activities {*Sit*, *Stand*, *Run*, and *Walk*}.

evolution of activities using a discrete-time, finite-state Markov chain (e.g., Fig. 8.9) defined on a given probability space $(\Omega, \mathcal{A}, P)$. The associated statistics are described by a $n \times n$ probability transition matrix $\mathbf{P}$, such that P_{jli} is the probability of the individual moving from state i to state j, that is, $P_{jli} = P(\mathbf{x}_{k+1} = \mathbf{e}_j \mid \mathbf{x}_k = \mathbf{e}_i)$ for $\mathbf{e}_i, \mathbf{e}_j \in \mathcal{X}$. We assume that the Markov chain is stationary, and thus these transition probabilities do not change with time. Estimation of the transition matrix $\mathbf{P}$ was beyond the scope of the current work. Note that approaches developed herein should extend to more complex, time-evolving activity alternation models.

At each time step, a set of biometric signals is produced by WBAN sensors. In a back-end server, a set of samples is generated by using appropriate feature extraction and selection techniques (see Refs. [18, 26]). A sample represents an extracted feature value from the generated biometric signals, for example, ECG period, fast Fourier transform of a window of biometric data, or ACC mean [18]. In Ref. [26], a filter-based feature selection method was proposed to determine the optimal feature set for the problem of physical activity detection. We assume that the generated samples come from this optimal feature set. To enable energy-efficient operation, the mobile phone chooses an appropriate control input at time k denoted by $\mathbf{u}_k$. Each control input is represented by a K-tuple of the form $\left[N_1^{\mathbf{u}_k}, N_2^{\mathbf{u}_k}, \ldots, N_K^{\mathbf{u}_k} \right]^{\mathrm{T}}$, where $N_l^{\mathbf{u}_k}$ denotes the total number of samples received from sensor S_l when control input $\mathbf{u}_k$ is selected. We assume that the total number of received samples from all selected sensors during the time interval $[k, k+1)$ satisfies the contraint $\mathbf{1}_K^{\mathrm{T}} \mathbf{u}_k \leqslant N$, where N is fixed. Based on the these definitions, we observe that $\mathbf{u}_k \in \mathcal{U} \triangleq \left\{ \mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^\alpha \right\}$, where $\alpha = \sum_{i=0}^{N} \binom{i+K-1}{i}$. We underscore that the state transitions are independent of the selected control input. We also observe that the number of control inputs is exponential in N and K, suggesting the need for low-cost approximation algorithms.

This description corresponds to a *discrete-time dynamical system* with system state $\mathbf{x}_k$ and control input $\mathbf{u}_k$. Since WBAN sensors communicate noisy measurements of the individual's physical activity, we are dealing with a *discrete-time dynamical system with imperfect or partially observed state information* (also referred to as a partially observable Markov decision process, or POMDP) [27].

Selecting a control input at time step $k-1$ results in a measurement vector $\mathbf{y}_k$ of the selected samples to be sent to the mobile phone. The selected samples follow an AR(1)-correlated multivariate Gaussian model, given that due to feature selection,

samples originating from the same and different sensors are assumed to be uncorrelated, but there exists temporal correlation for a single sample [26]. As a result, we have the following n-ary generalized Gaussian hypothesis testing problem

$$H_i^{\mathbf{u}_{k-1}} : \mathbf{y}_k \sim \mathcal{N}\left(\mathbf{m}_{i,\mathbf{u}_{k-1}}, \mathbf{Q}_{i,\mathbf{u}_{k-1}}\right), \quad \mathbf{e}_i \in \mathcal{X}, \tag{8.1}$$

where $\mathbf{m}_{i,\mathbf{u}_{k-1}}, \mathbf{Q}_{i,\mathbf{u}_{k-1}}$ represent the mean vectors and covariance matrices of the measurement vector $\mathbf{y}_k$ under each hypothesis, given control $\mathbf{u}_{k-1}$. The mean vectors and covariance matrices are defined as follows:

$$\mathbf{m}_{i,\mathbf{u}_{k-1}} \triangleq \left[\boldsymbol{\mu}_{i,\mathbf{u}_{k-1}}\left(S_1\right)^{\mathrm{T}}, \boldsymbol{\mu}_{i,\mathbf{u}_{k-1}}\left(S_2\right)^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_{i,\mathbf{u}_{k-1}}\left(S_K\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{8.2}$$

$$\mathbf{Q}_{i,\mathbf{u}_{k-1}} \triangleq \operatorname{diag}\left(\boldsymbol{\mathcal{Q}}_{i,\mathbf{u}_{k-1}}\left(S_1\right), \boldsymbol{\mathcal{Q}}_{i,\mathbf{u}_{k-1}}\left(S_2\right), \ldots, \boldsymbol{\mathcal{Q}}_{i,\mathbf{u}_{k-1}}\left(S_K\right)\right). \tag{8.3}$$

Here, $\boldsymbol{\mu}_{i,\mathbf{u}_{k-1}}\left(S_l\right)$ is a $N_l^{\mathbf{u}_{k-1}} \times 1$ vector, $\boldsymbol{\mathcal{Q}}_{i,\mathbf{u}_{k-1}}\left(S_l\right) = \left(\sigma_{S_l,i}^2 / 1-\phi^2\right)\mathbf{T} + \sigma_z^2\mathbf{I}$ is a $N_l^{\mathbf{u}_{k-1}} \times N_l^{\mathbf{u}_{k-1}}$ matrix, $\mathbf{T}$ is a Toeplitz matrix whose first row/column is $\left[1, \phi, \phi^2, \ldots, \phi^{N_l^{\mathbf{u}_{k-1}}-1}\right]$, ϕ is the parameter of the AR(1) model, and σ_z^2 accounts for sensing and communication noise. We observe that assuming two distinct control inputs $\mathbf{u}^{q_1}$ and $\mathbf{u}^{q_2}, q_1, q_2 \in \mathcal{U}$ and a WBAN sensor S_l, the associated covariance matrices $\boldsymbol{\mathcal{Q}}_{i,\mathbf{u}^{q_1}}(S_l)$ and $\boldsymbol{\mathcal{Q}}_{i,\mathbf{u}^{q_2}}(S_l)$ are of different dimensions. In addition, the size of $\mathbf{Q}_{i,\mathbf{u}_{k-1}}$ is determined by the total number of received samples at each time step.

At the mobile phone, the received measurement vector $\mathbf{y}_k$ is passed through an instantaneous detector structure, which produces an estimate of the activity state at time k, denoted by $\mathbf{z}_k \in \mathcal{Z} \triangleq \mathcal{X} \cup \{\varepsilon\}$. The erasure value ε corresponds to the case, where the control input $\mathbf{0}_K$ is selected, in which case no measurements are generated. For the n-ary problem (8.1), there is no closed-form expression for the observation probabilities $P\left(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{u}_{k-1}\right)$, where $\mathbf{u}_{k-1} \neq \mathbf{0}_K$. Therefore, we use the following pairwise error probability upper bound [28] to approximate them

$$r\left(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, \mathbf{x}_k, \mathbf{z}_k\right) \triangleq P\left(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{u}_{k-1}\right) \cong \sqrt{P_{\mathbf{z}_k|\mathbf{x}_{k-1}} P_{\mathbf{x}_k|\mathbf{x}_{k-1}}}\, \rho_{\mathbf{z}_k,\mathbf{x}_k}^{\mathbf{u}_{k-1}}, \tag{8.4}$$

where $\rho_{\mathbf{z}_k,\mathbf{x}_k}^{\mathbf{u}_{k-1}}$ is the Bhattacharyya coefficient [29] that in the case of multivariate Gaussian random vectors takes the form

$$\begin{aligned}
&\rho\left(\mathbf{z}_k, \mathbf{z}_k, \mathbf{u}_{k-1}\right) \\
&= \exp\left(-\left[\frac{1}{8}\left(\Delta\mathbf{m}_{\mathbf{z}_k,\mathbf{x}_k}^{\mathbf{u}_{k-1}}\right)^{\mathrm{T}}\left(\mathbf{Q}_h^{\mathbf{u}_{k-1}}\right)^{-1}\Delta\mathbf{m}_{\mathbf{z}_k,\mathbf{x}_k}^{\mathbf{u}_{k-1}} + \frac{1}{2}\log\frac{\det\mathbf{Q}_h^{\mathbf{u}_{k-1}}}{\sqrt{\det\mathbf{Q}_{\mathbf{x}_k}^{\mathbf{u}_{k-1}} \cdot \det\mathbf{Q}_{\mathbf{z}_k}^{\mathbf{u}_{k-1}}}}\right]\right),
\end{aligned} \tag{8.5}$$

with $\Delta\mathbf{m}_{\mathbf{z}_k,\mathbf{x}_k}^{\mathbf{u}_{k-1}} = \left(\mathbf{m}_{\mathbf{z}_k}^{\mathbf{u}_{k-1}} - \mathbf{m}_{\mathbf{x}_k}^{\mathbf{u}_{k-1}}\right)$ and $2\mathbf{Q}_h^{\mathbf{u}_{k-1}} = \mathbf{Q}_{\mathbf{z}_k}^{\mathbf{u}_{k-1}} + \mathbf{Q}_{\mathbf{x}_k}^{\mathbf{u}_{k-1}}$. Note that the approximation in (8.4) introduces a dependence on the physical activity performed by the individual in the previous time step, which can be explained by the Markovian nature of the state process. Computing the observation probabilities can be accomplished off-line and requires $\mathcal{O}\left(n^3\alpha N^3\right)$.

The total information available for decision making at the mobile phone at time step k corresponds to the *information vector* $I_k \triangleq (\mathbf{z}_0,\ldots,\mathbf{z}_k,\mathbf{u}_0,\ldots,\mathbf{u}_{k-1})$, where $I_0 \triangleq \mathbf{z}_0$ [27]. The control input at time step k is selected based on this information vector, that is, $\mathbf{u}_k = \eta_k(I_k)$ with $\eta_k(\cdot)$ denoting the sensor selection strategy at time step k. Next, we present the performance measures we employ.

8.4.2 Performance Measures

Sensor heterogeneity is characterized by *detection accuracy* and *energy consumption*. In particular, we use the *worst-case error probability* $p_e^{\mathcal{W}}(\mathbf{x}_k,\mathbf{u}_k)$ to characterize the discriminative capabilities of the heterogeneous sensors by exploiting the fact that different control inputs lead to different detection accuracy. This is defined as

$$p_e^{\mathcal{W}}(\mathbf{x}_k,\mathbf{u}_k) \triangleq \max_{\substack{\mathbf{x}_{k+1},\mathbf{z}_{k+1} \\ \mathbf{x}_{k+1}\neq\mathbf{z}_{k+1}}} \left[r(\mathbf{x}_k,\mathbf{u}_k,\mathbf{x}_{k+1},\mathbf{z}_{k+1}) \right], \tag{8.6}$$

where $p_e^{\mathcal{W}}(\mathbf{x}_k,\mathbf{u}_k) \in [0,1]$, $\forall \mathbf{x}_k \in \mathcal{X}$, $\mathbf{u}_k \in \mathcal{U}$, and $p_e^{\mathcal{W}}(\mathbf{x}_k,\mathbf{0}_K) = 1$, $\forall \mathbf{x}_k \in \mathcal{X}$.

To capture the energy consumption, we use the *unnormalized energy cost* of control input $\mathbf{u}_k$ defined as $e(\mathbf{u}_k) \triangleq \mathbf{u}_k^{\mathrm{T}}\boldsymbol{\delta}$, where $\boldsymbol{\delta} \triangleq [\delta_1,\ldots,\delta_K]^{\mathrm{T}}$. The notation $\delta_l \in (0,1]$ indicates the reception cost of one sample from the sensor S_l, which is fixed and known for each sensor, as we saw earlier in Section 8.2. As a reminder, different WBAN sensors exhibit different energy costs because of their distinct data rates and location (internal versus external sensors). We also define the *normalized energy cost* $\mathcal{E}(\mathbf{u}_k) \in [0,1]$. Both costs do not depend on the individual's physical activity, but they depend on the selected sensors and the total and per sensor number of samples.

To study the trade-off between detection accuracy and energy consumption, we define the *total cost*

$$g^{\lambda}(\mathbf{x}_k,\mathbf{u}_k) \doteq (1-\lambda)\, p_e^{\mathcal{W}}(\mathbf{x}_k,\mathbf{u}_k) + \lambda \mathcal{E}(\mathbf{u}_k), \tag{8.7}$$

where $\lambda \in [0,1]$ captures the relative importance between worst-case error probability and normalized energy cost.

8.4.3 Partial Observability and Sufficient Statistics

To overcome the complications in decision making introduced by the expanding dimension of the information vector I_k at each time step, we use the *belief state* [27] as a sufficient statistic. This corresponds to the probability distribution of state $\mathbf{x}_k$ given I_k and is defined as

$$\mathbf{p}_k = \left[p_k^1, p_k^2, \ldots, p_k^n \right]^{\mathrm{T}} \in \mathcal{P}, \quad \text{where } p_k^j = P(\mathbf{x}_k = \mathbf{e}_j \mid I_k),\ j \in \mathcal{X}, \tag{8.8}$$

where $\mathcal{P}$ denotes the *belief space*, which is a $n-1$ dimensional simplex defined as follows:

$$\mathcal{P} = \left\{ \mathbf{p}_k \in \mathbb{R}^n : \mathbf{1}_n^{\mathrm{T}}\mathbf{p}_k = 1, p_k^j \in [0,1], \forall j \in \mathcal{X} \right\}. \tag{8.9}$$

The belief state time evolution is accomplished via the update rule given in Lemma 8.1.

Lemma 8.1 Zois *et al.* [2]
Let $\mathbf{p}_k$ *denote the belief state at time step k. Assume that the control input* $\mathbf{u}_k$ *is selected and at time step* $k+1$*, the observation* $\mathbf{z}_{k+1}$ *is generated. Then, the belief state* $\mathbf{p}_{k+1}$ *is given by the following rule:*

$$\mathbf{p}_{k+1} = \frac{\left[\mathbf{P} \circ \mathbf{r}\left(\mathbf{u}_k, \mathbf{z}_{k+1}\right)\right]^{\mathrm{T}} \mathbf{p}_k}{\mathbf{1}_n^{\mathrm{T}} \left[\mathbf{P} \circ \mathbf{r}\left(\mathbf{u}_k, \mathbf{z}_{k+1}\right)\right]^{\mathrm{T}} \mathbf{p}_k}. \tag{8.10}$$

Here, $\circ$ *denotes the Hadamard product between the matrices of interest and* $\mathbf{r}\left(\mathbf{u}_k, \mathbf{z}_{k+1}\right)$ *denotes the* $n \times n$ *matrix of observation probabilities with elements* $r\left(i, \mathbf{u}_k, j, \mathbf{z}_{k+1}\right)$ *and* $i, j \in \mathcal{X}$.

We observe that the aforementioned update rule involves the Hadamard product of two matrices, resulting in a more complex evolution compared to standard POMDP models, further challenging the solution of the optimization problem.

In the case where control input $\mathbf{0}_K$ is selected, the update rule of Lemma 8.1 cannot be used. Instead, we take advantage of the Markovian nature of the physical activity process and use $\mathbf{p}_{k+1} = \mathbf{P}\mathbf{p}_k$. In any case, we can estimate the individual's physical activity by employing a maximum likelihood (ML) detector of the form $\hat{\mathbf{x}}_k^{\mathrm{ML}} = \arg\max \mathbf{p}_k$.

8.4.4 Optimization Problem

We define the following cost function:

$$J^{\lambda} = \mathbb{E}\left\{\sum_{k=0}^{L-1} g^{\lambda}\left(\mathbf{x}_k, \mathbf{u}_k\right)\right\}. \tag{8.11}$$

Here, $L < \infty$ denotes the horizon length, and the terminal cost $g(\mathbf{x}_L)$ is assumed to be zero. Our aim is to find the optimal sensor selection strategy that minimizes the total accumulated cost J^{λ} from time 1 to time L over the set of admissible control policies. This corresponds to the following *finite-horizon, partially observable stochastic control problem*: $\min_{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{L-1}} J^{\lambda}$. We observe that the worst-case error probability and the normalized energy cost are uniformly bounded from below and above. In addition, the state and control spaces are finite sets. Thus, according to Ref. [27], an optimal sensor selection strategy exists and in fact, there is a different optimal sensor selection strategy for each distinct value of λ.

8.5 SENSOR SELECTION STRATEGIES

8.5.1 Optimal Sensor Selection

We begin by presenting the stochastic dynamic programming (DP) algorithm that determines the optimal sensor selection strategy. Contrary to traditional POMDP models, our model is characterized by two unique features. First, the control input

specifies the number of received observations in each step and their associated quality, but does not influence the physical activity transitions. Second, observation probabilties depend on the current and previous physical activity as well as the control input. Theorem 8.1 provides DP equations that determine the optimal sensor selection strategy in this context.

Theorem 8.1 Zois *et al.* [2]
For $k = L-2,\ldots,0$, the cost-to-go function $\overline{J}_k^{\lambda}(\mathbf{p}_k)$ is related to $\overline{J}_{k+1}^{\lambda}(\mathbf{p}_{k+1})$ through the recursion

$$\overline{J}_k^{\lambda}(\mathbf{p}_k) = \min_{\mathbf{u}_k \in \mathcal{U}} \left[\mathbf{p}_k^{\mathrm{T}} \mathbf{g}^{\lambda}(\mathbf{u}_k) + \mathcal{A}(\mathbf{p}_k, \mathbf{u}_k) \right], \tag{8.12}$$

where

$$\mathcal{A}(\mathbf{p}_k, \mathbf{u}_k) = \begin{cases} \sum_{\theta=1}^{n} \mathbf{1}_n^{\mathrm{T}} \left[\mathbf{P} \circ \mathbf{r}(\mathbf{u}_k, \mathbf{e}_\theta) \right]^{\mathrm{T}} \mathbf{p}_k \overline{J}_{k+1}^{\lambda} \left(\dfrac{\left[\mathbf{P} \circ \mathbf{r}(\mathbf{u}_k, \mathbf{e}_\theta) \right]^{\mathrm{T}} \mathbf{p}_k}{\mathbf{1}_n^{\mathrm{T}} \left[\mathbf{P} \circ \mathbf{r}(\mathbf{u}_k, \mathbf{e}_\theta) \right]^{\mathrm{T}} \mathbf{p}_k} \right), & \mathbf{u}_k \neq \mathbf{0}_K \\ \overline{J}_{k+1}^{\lambda}(\mathbf{P}\mathbf{p}_k), & \mathbf{u}_k = \mathbf{0}_K \end{cases}. \tag{8.13}$$

The cost-to-go function for $k = L-1$ is given by

$$\overline{J}_{L-1}^{\lambda}(\mathbf{p}_{L-1}) = \min_{\mathbf{u}_{L-1} \in \mathcal{U}} \left[\mathbf{p}_{L-1}^{\mathrm{T}} \mathbf{g}^{\lambda}(\mathbf{u}_{L-1}) \right]. \tag{8.14}$$

If $\mathbf{u}_k^* = \eta_{k+1}^*(\mathbf{p}_k)$ minimizes the right-hand side of DP for each k, the optimal sensor selection strategy will be $\eta^* = \{\eta_0^*, \eta_1^*, \ldots, \eta_{L-1}^*\}$. As in traditional POMDP models, our POMDP has an uncountably infinite belief space $\mathcal{P}$. Furthermore, the control space $\mathcal{U}$ is exponentially large, and the observation probabilities also depend on the previous physical activity. The last two facts exacerbate DP's already expensive computational complexity. In particular, the computational complexity of finding the optimal sensor selection strategy via DP in our case is $\mathcal{O}(n^3 (d+1)^n \alpha L)$, where d denotes the quantization resolution of the belief space $\mathcal{P}$.

8.5.2 Suboptimal Sensor Selection

We have already asserted that using DP to compute the optimal sensor selection strategy is a considerably expensive task in terms of computation. Next, we propose alternative, low-cost suboptimal solutions that can be implemented and run in real time.

8.5.2.1 MIC-T3S Algorithm To design a practical algorithm, we will exploit a number of properties of the cost-to-go function $\overline{J}_k^{\lambda}(\mathbf{p}_k)$ that we prove. We start with Lemma 8.2 that can be shown by induction.

Lemma 8.2 Zois *et al.* [2]
The function $\overline{J}_k^{\lambda}\left(\mathbf{p}_k\right)$, $k = L-1, L-2, \ldots, 0$, *is positively homogeneous of degree* 1 *that is,*

$$\overline{J}_k^{\mu}\left(\mu\mathbf{p}_k\right) = \mu\overline{J}_k^{\mu}\left(\mathbf{p}_k\right), \quad \forall\mu > 0. \tag{8.15}$$

Next, we observe that the term $\mathbf{1}_n^{\mathrm{T}}\left[\mathbf{P}\circ\mathbf{r}\left(\mathbf{u}_k,\mathbf{e}_\theta\right)\right]^{\mathrm{T}}\mathbf{p}_k$ is a scalar for any control input $\mathbf{u}_k \in \mathcal{U}\setminus\mathbf{0}_K$. Therefore, we can simplify the term $\mathcal{A}\left(\mathbf{p}_k,\mathbf{u}_k\right)$, $\forall\mathbf{u}_k \in \mathcal{U}\setminus\mathbf{0}_K$ in (8.13) by setting $\mu = \mathbf{1}_n^{\mathrm{T}}\left[\mathbf{P}\circ\mathbf{r}\left(\mathbf{u}_k,\mathbf{e}_\theta\right)\right]^{\mathrm{T}}\mathbf{p}_k$ and exploiting Lemma 8.2.

Lemma 8.3 Zois *et al.* [2]
The function $\mathcal{A}\left(\mathbf{p}_k,\mathbf{u}_k\right)$ *can be written in the following simpler form:*

$$\mathcal{A}\left(\mathbf{p}_k,\mathbf{u}_k\right) = \begin{cases} \sum_{\theta=1}^{n}\overline{J}_{k+1}^{\lambda}\left(\left[\mathbf{P}\circ\mathbf{r}\left(\mathbf{u}_k,\mathbf{e}_\theta\right)\right]^{\mathrm{T}}\mathbf{p}_k\right), & \mathbf{u}_k \neq \mathbf{0}_K \\ \overline{J}_{k+1}^{\lambda}\left(\mathbf{P}\mathbf{p}_k\right), & \mathbf{u}_k = \mathbf{0}_K \end{cases} \tag{8.16}$$

for $k = L-2, \ldots, 0$.

Lemma 8.3 and DP equations in (8.12)–(8.14) enable us to prove Lemma 8.4 by induction.

Lemma 8.4
The function $\overline{J}_k^{\lambda}\left(\mathbf{p}_k\right)$, $k = L-1, \ldots, 0$, *is concave and piece-wise linear.*

Next, we state Theorem 8.2.

Theorem 8.2 Rockaffeller and Wets [30]
A function $\Gamma : \mathbb{R}^a \to \mathbb{R}^b$ *is superlinear if and only if* Γ *is positively homogeneous and concave. Moreover, then* $\Gamma\left(\sum_{i=1}^{n} f^i\right) \geq \sum_{i=1}^{n}\Gamma\left(f^i\right)$.

We are now ready to describe the main idea behind our proposed suboptimal algorithm. In particular, the cost-to-go function $\overline{J}_k^{\lambda}\left(\mathbf{p}_k\right)$ is positively homogeneous and concave, as verified by Lemmas 8.2 and 8.4. In addition, according to Theorem 8.2, $\overline{J}_k^{\lambda}\left(\mathbf{p}_k\right)$ is also superlinear. Therefore, if we express any arbitrary belief state $\mathbf{p}_k$ as

$$\mathbf{p}_k = p_k^1\mathbf{e}_1 + \cdots + p_k^n\mathbf{e}_n, \tag{8.17}$$

then we have that

$$\overline{J}_k^{\lambda}\left(\mathbf{p}_k\right) \geq p_k^1\overline{J}_k^{\lambda}\left(\mathbf{e}_1\right) + \cdots + p_k^n\overline{J}_k^{\lambda}\left(\mathbf{e}_n\right). \tag{8.18}$$

If we optimize each individual term on the right-hand side of (8.18), we get the optimal sensor selection strategy for the states in the corners of the belief space $\mathcal{P}$, weighted by the probability of being in these states. We can exploit (8.18) in conjunction with the fact that each control input $\mathbf{u}_k$ is a vector comprising of the number of received samples per sensor to determine the control input at an arbitrary belief state $\mathbf{p}_k$ by time-sharing as $\hat{\mathbf{u}}_k^{\mathbf{p}_k} = p_k^1\hat{\mathbf{u}}_k^{\mathbf{e}_1} + \cdots + p_k^n\hat{\mathbf{u}}_k^{\mathbf{e}_n}$. In the case where noninteger

values arise, we enforce a minimum distance rule where $\hat{\mathbf{u}}_k^{\mathbf{p}_k}$ is assigned the closest control. Ties are broken arbitrarily.

At this point, we exploit the right-hand side of (8.18) and along with (8.12) and (8.16), we get the following approximation:

$$\hat{J}_k^{\lambda}(\mathbf{p}_k) \approx \min_{\mathbf{u}_k \in \mathcal{U}} \Bigg[\mathbf{p}_k^{\mathrm{T}} \mathbf{g}^{\lambda}(\mathbf{u}_k) + \sum_{\theta=1}^{n} \sum_{i=1}^{n} \Big[\Big[\mathbf{P} \circ \mathbf{r}(\mathbf{u}_k, \mathbf{e}_\theta) \Big]^{\mathrm{T}} \mathbf{p}_k \Big]_i \hat{J}_{k+1}^{\lambda}(\mathbf{e}_i), \\ \mathbf{p}_k^{\mathrm{T}} \mathbf{g}^{\lambda}(\mathbf{0}_K) + \sum_{i=1}^{n} \big[\mathbf{P} \mathbf{p}_k \big]_i \hat{J}_{k+1}^{\lambda}(\mathbf{e}_i) \Bigg], \quad k = L-2, \ldots, 0. \tag{8.19}$$

$[]_i$ denotes the i-th element. We propose the minimum integrated cost time-sharing sensor selection (MIC-T3S) algorithm, which entails the execution of the following two components:

1. *Offline*: We run Algorithm 8.1 at a remote back-end server to determine the sensor selection strategy at the corners of the belief space $\mathcal{P}$ for $k = 0, 1, \ldots, L-1$. The associated time and space complexity is $\mathcal{O}(n^4 \alpha L)$ and $\mathcal{O}(n^3 \alpha)$, respectively.
2. *Online*: During the system's normal operation, we employ time-sharing along with the update rule for $\mathbf{p}_k$ in (8.10). The associated time and space complexity is $\mathcal{O}(n^4)$ and $\mathcal{O}(n^3 \alpha)$, respectively.

Algorithm 8.1 MIC-T3S (*Offline part*)

Input: transition matrix $\mathbf{P}$, observation matrix $\mathbf{r}(\mathbf{u}_k, z_{k+1})$, costs $\mathbf{g}^{\lambda}(\mathbf{u}^q), q = 1, \ldots, \alpha$, and horizon length L

Output: cost $\left\{ \hat{J}_k^{\lambda}(\mathbf{e}_1), \ldots, \hat{J}_{L-1}^{\lambda}(\mathbf{e}_n) \right\}_{k=0}^{L-1}$ and strategy $\left\{ \hat{\mathbf{u}}_k^{\mathbf{e}_1}, ,, \ldots, , \hat{\mathbf{u}}_k^{\mathbf{e}_n} \right\}_{k=0}^{L-1}$

1: **for** $i = 1 : n$ **do**
2: $\quad \hat{J}_{L-1}^{\lambda}(\mathbf{e}_i) = \min_{\mathbf{u}_{L-1} \in \mathcal{U}} \left[\mathbf{e}_i^{\mathrm{T}} \mathbf{g}^{\lambda}(\mathbf{u}_{L-1}) \right]$;
3: $\quad \hat{\mathbf{u}}_{L-1}^{\mathbf{e}_i} = \arg\min_{\mathbf{u}_{L-1} \in \mathcal{U}} \hat{J}_{L-1}^{\lambda}(\mathbf{e}_i)$;
4: **end for**
5: **for** $k = L-2 : 0$ **do**
6: $\quad$ **for** $i = 1 : n$ **do**
7: $\quad\quad$ Use (8.19) to determine $\hat{J}_k^{\lambda}(\mathbf{e}_i)$ and $\hat{\mathbf{u}}_k^{\mathbf{e}_i}$;
8: $\quad$ **end for**
9: **end for**

8.5.2.2 Constrained Algorithms In contrast to MIC-T3S, which focuses on the trade-off between detection accuracy and energy consumption, herein we consider the cost function (i.e., $\lambda = 1$):

$$J^1 = \mathbb{E}\left\{ \sum_{k=0}^{L-1} g^1(x_k, \mathbf{u}_k) \right\} = \mathbb{E}\left\{ \sum_{k=0}^{L-1} \mathcal{E}(\mathbf{u}_k) \right\}. \tag{8.20}$$

Clearly, if we do not impose any constraints during the minimization of J^1, the resulting sensor selection strategy will have minimum energy consumption but probably very poor detection accuracy. To overcome this issue, we introduce two instantaneous constraints that lead to two different suboptimal strategies.

1. *E²MBADP algorithm*: We wish to minimize (8.20) and at the same time choose control inputs that will result in high certainty belief states. This results in the following optimization problem:

$$\min_{\mathbf{u}_0,\ldots,\mathbf{u}_{L-1}} J^1 \quad \text{subject to } \max\left(\mathbf{p}_k\right) \geq \tau. \tag{8.21}$$

Here, $\tau \in [0,1]$. The main idea of the energy-efficient maximal belief approximate DP (E²MBADP) shown in algorithm is to use DP and prune infeasible control inputs as we go based on the constraint in (8.21). In the case where the constraint is not satisfied, we choose the control input that results in detection accuracy closest to threshold τ.

Algorithm 8.2 E²MBADP

Input: set of belief states $\{\mathbf{p}_1,\ldots,\mathbf{p}_{bf}\}$, transition matrix $\mathbf{P}$, observation matrix $\mathbf{r}\left(\mathbf{u}_k, z_{k+1}\right)$, costs $\mathcal{E}\left(\mathbf{u}^q\right), q = 1,\ldots,\alpha$, and horizon length L

Output: cost $\left\{\bar{J}_k^1\left(\mathbf{p}_1\right),\ldots,\bar{J}_k^1\left(\mathbf{p}_{bf}\right)\right\}_{k=0}^{L-1}$ and strategy $\left\{\mathbf{u}_k^{\mathbf{p}_1},,\ldots,,\mathbf{u}_k^{\mathbf{p}_{bf}}\right\}_{k=0}^{L-1}$

1: **for** $i = 1 : bf$ **do**
2: $\quad \bar{J}_{L-1}^1\left(\mathbf{p}_i\right) = \min_{\mathbf{u}_{L-1}\in\mathcal{U}}\left[\mathbf{p}_i^{\mathrm{T}}\mathcal{E}\left(\mathbf{u}_{L-1}\right)\right]$ subject to $\max\left(\mathbf{p}_i\right) \geq \tau$;
3: $\quad \mathbf{u}_{L-1}^{\mathbf{p}_i} = \arg\min_{\mathbf{u}_{L-1}\in\mathcal{U}} \bar{J}_{L-1}^1\left(\mathbf{p}_i\right)$ subject to $\max\left(\mathbf{p}_i\right) \geq \tau$;
4: **end for**
5: **for** $k = L-2 : 0$ **do**
6: $\quad$ **for** $i = 1 : bf$ **do**
7: $\quad\quad \bar{J}_k^1\left(\mathbf{p}_i\right) = \min_{\mathbf{u}_k\in\mathcal{U}}\left[\mathbf{p}_i^{\mathrm{T}}\mathcal{E}\left(\mathbf{u}_k\right) + \mathcal{A}\left(\mathbf{p}_i,\mathbf{u}_k\right)\right]$ subject to $\max\left(\mathbf{p}_i\right) \geq \tau$;
8: $\quad\quad \mathbf{u}_k^{\mathbf{p}_i} = \arg\min_{\mathbf{u}_k\in\mathcal{U}} \bar{J}_k^1\left(\mathbf{p}_i\right)$ subject to $\max\left(\mathbf{p}_i\right) \geq \tau$;
9: $\quad$ **end for**
10: **end for**

2. *GME²PS² algorithm*: We wish to minimize (8.20) and at the same time choose control inputs that will result in low worst-case error probabilities in the next time step. This results in the following optimization problem:

$$\min_{\mathbf{u}_0,\ldots,\mathbf{u}_{L-1}} J^1 \quad \text{subject to } P_e^{\mathcal{W}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \leq 1-\tau. \tag{8.22}$$

Here, τ is defined as before. To solve this, we propose the greedy minimum energy and error probability sensor selection (GME²PS²) algorithm. At each

time step, the unknown physical activity is estimated using the ML detector on the incoming belief state $\mathbf{p}_k$. Next, we choose the control input $\mathbf{u}_k$ that minimizes the instantaneous energy cost $\varepsilon(\mathbf{u}_k)$ subject to $P_e^{\mathcal{W}}\left(\hat{\mathbf{x}}_k^{\mathrm{ML}}, \mathbf{u}_k\right) \leq 1-\tau$.

Both E^{2}MBADP and GME^{2}PS^2 wish to minimize the energy consumption, while satisfying an accuracy constraint. In case this cannot be achieved, they select the "best of the worst" control inputs, that is, the one that leads to accuracy closest to the desired threshold. Their key difference lies on the constraint definition in conjunction with the steps they perform to satisfy it. Namely, E^{2}MBADP chooses control inputs, which will result in accurate belief state estimates while ensuring that energy consumption is minimum. GME^{2}PS^2, on the other hand, passes the incoming belief state through the ML detector to receive a physical activity estimate and then uses this estimate to drive control input selection. We expect that for low values of τ, E^{2}MBADP will behave slightly worse than GME^{2}PS^2, since in the case of the former algorithm, the constraint in (8.21) may be satisfied but the rest of the states may have nonzero belief values. If these physical activities can be easily confused with the true physical activity, more detection errors will occur. To avoid this issue, GME^{2}PS^2 chooses cleverly between available control inputs to also bound the worst-case error probability.

8.6 ALTERNATIVE PROBLEM FORMULATION

In the problem formulation introduced in Section 8.4, we have assumed the existence of an instantaneous detector that produces physical activity estimates based on the received measurement vectors. In this section, exploiting the knowledge we have gained on the design of sensor selection strategies, we work with a more realistic but also more complicated model, where the mobile phone directly uses the received measurement vector. In other words, we adopt the system model of Section 8.4 and discard the instantaneous detector. Once more, our goal is to accurately infer the unknown, time-evolving physical activity of the individual, while ensuring the energy-efficient operation of the WBAN system. To this end, we address the problem of determining expressions for the minimum mean square error (MMSE) system state estimate and the optimal sensor selection strategy that optimizes the trade-off between the estimator's performance and the system's energy consumption.

8.6.1 Kalman-Like Estimator

As already discussed in Section 8.4.3, the belief state $\mathbf{p}_k$ constitutes an estimate of the physical activity at each time step and drives the sensor selection process. In fact, the belief state $\mathbf{p}_k$ coincides with the optimal MMSE state estimate of the physical activity, since

$$\hat{\mathbf{x}}_{\mathrm{MMSE}} \triangleq \mathbb{E}\left\{\mathbf{x}_k \mid \mathcal{F}_k\right\} = \sum_{i=1}^{n} \mathbf{e}_i P\left(\mathbf{x}_k = \mathbf{e}_i \mid I_k\right) = \mathbf{p}_k, \tag{8.23}$$

where now $I_k \triangleq \left(\mathbf{y}_0, \ldots, \mathbf{y}_k, \mathbf{u}_0, \ldots, \mathbf{u}_{k-1}\right)$ and $I_0 \triangleq \mathbf{y}_0$.

One of the major contributions in modern control theory is the renowned Kalman filter [31]. It is the optimal MMSE estimator in the linear quadratic Gaussian case and has a very appealing recursive form. Namely, the MMSE estimate and the associated MSE are given by the conditional mean and covariance matrix, respectively, of a Gaussian multivariate model, and are recursively updated in an efficient way when a new measurement arrives. This particular structure has enabled the application of the Kalman filter in diverse domains, such as signal processing, communications, economics, neuroscience, machine learning, computer vision, aerospace, and political science. Inspired by the success of the Kalman filter, our goal is to derive, if possible, recursive expressions for the optimal MMSE physical activity estimator in (8.23) requiring the Kalman filter structure. *Unfortunately, for our system model, the optimal Kalman-like MMSE estimator is intrinsically nonrecursive* [32]. Since we desire a recursive solution within the family of Kalman-like estimators, we impose recursivity as a design constraint and derive an approximate MMSE estimator. Theorem 8.3 provides the recursive expressions for this estimator denoted hereafter by $\hat{\mathbf{p}}_k$.

Theorem 8.3 Zois *et al.* [32]
The Markov chain system estimate at time step k is recursively defined as

$$\hat{\mathbf{p}}_k = \hat{\mathbf{p}}_k^- + \mathbf{G}_k\left(\mathbf{y}_k - \mathbf{y}_k^-\right), k \geq 0 \tag{8.24}$$

with

$$\hat{\mathbf{p}}_k^- = \mathbf{P}\hat{\mathbf{p}}_{k-1}, \tag{8.25}$$

$$\mathbf{y}_k^- = \mathcal{M}\left(\mathbf{u}_{k-1}\right)\hat{\mathbf{p}}_k^-, \tag{8.26}$$

$$\mathbf{G}_k = \sum\nolimits_k^- \mathcal{M}^{\mathrm{T}}\left(\mathbf{u}_{k-1}\right)\left(\mathcal{M}\left(\mathbf{u}_{k-1}\right)\sum\nolimits_k^- \mathcal{M}^{\mathrm{T}}\left(\mathbf{u}_{k-1}\right) + \tilde{\mathbf{Q}}_k\right)^{-1}, \tag{8.27}$$

where $\hat{\mathbf{p}}_0^- = \pi$, *and* π *is the initial distribution over the system states,* $\mathcal{M}\left(\mathbf{u}_{k-1}\right) \triangleq \left[\mathbf{m}_{1,\mathbf{u}_{k-1}}, \ldots, \mathbf{m}_{n,\mathbf{u}_{k-1}}\right]$, $\sum_k^-$ *is the conditional covariance matrix of the prediction error, and* $\tilde{\mathbf{Q}}_k \triangleq \sum_{i=1}^{n} \hat{p}_k^{-,i}\mathbf{Q}_{i,\mathbf{u}_{k-1}}$.

The proposed estimator is *formally* similar to the standard Kalman filter [31], but its gain $\mathbf{G}_k$ *depends on the measurement vectors*. As a result, it constitutes a nonlinear estimator versus the standard linear Kalman filter (see Ref. [32] for a detailed comparison between the two estimators). Since we do not impose any constraints on the components of $\hat{\mathbf{p}}_k$, we cannot guarantee that they will lie in the [0, 1] interval. We apply an appropriate transformation to $\hat{\mathbf{p}}_k$ to ensure feasible solutions are determined. In particular, we adopt the following transformation:

$$\mathcal{T}\left(\hat{\mathbf{p}}_k\right) = \frac{\mathbf{A}\hat{\mathbf{p}}_k}{\mathbf{1}_n^{\mathrm{T}}\mathbf{A}\hat{\mathbf{p}}_k} \quad \text{with } a_{ii} = \mathbf{1}_{\left\{\hat{p}_k^i \geq 0\right\}} \text{ and } a_{ij} = 0,\ \forall i \neq j. \tag{8.28}$$

The adopted transformation is nonlinear, data dependent, and assigns negative elements to zeros while ensuring that the end result will sum to 1. The interested reader is referred to Ref. [32] to learn more about this transformation.

The MSE performance of the estimator in (8.24) is captured by the *conditional filtering error covariance matrix*

$$\sum\nolimits_k \triangleq \mathbb{E}\left\{\left(\mathbf{x}_k - \hat{\mathbf{p}}_k\right)\left(\mathbf{x}_k - \hat{\mathbf{p}}_k\right)^{\mathrm{T}} \mid \mathcal{F}_k\right\} = \mathrm{diag}\left(\hat{\mathbf{p}}_k\right) - \hat{\mathbf{p}}_k\hat{\mathbf{p}}_k^{\mathrm{T}}. \tag{8.29}$$

Likewise, the MSE performance of the predictor in (8.25) is given by the *conditional prediction error covariance matrix*

$$\sum\nolimits_k^- \triangleq \mathbb{E}\left\{\left(\mathbf{x}_k - \hat{\mathbf{p}}_k^-\right)\left(\mathbf{x}_k - \hat{\mathbf{p}}_k^-\right)^{\mathrm{T}} \mid \mathcal{F}_{k-1}\right\} = \mathrm{diag}\left(\hat{\mathbf{p}}_k^-\right) - \hat{\mathbf{p}}_k^-\hat{\mathbf{p}}_k^{-,\mathrm{T}}. \tag{8.30}$$

8.6.2 Optimization Problem

We have already pointed out that we wish to infer the time-evolving physical activity of the individual but also guarantee the energy-efficient operation of the WBAN system. Thus, we are interested in two performance metrics: the *estimation accuracy* and the *energy consumption*. For the latter metric, we use the metrics introduced in Section 8.6.3, that is, the unnormalized energy cost $e(\mathbf{u}_k)$ and the normalized energy cost $\varepsilon(\mathbf{u}_k)$. On the other hand, we use the *estimation accuracy* $\mathrm{MSE}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right)$ to capture the discriminative capabilities of the heterogeneous sensors defined as follows:

$$\mathrm{MSE}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \triangleq \mathrm{tr}\left(\sum_k \left(\mathbf{y}_k, \mathbf{u}_{k-1}\right)\right). \tag{8.31}$$

Here, $\mathrm{MSE}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \in [0,1]$. Note that the estimation accuracy metric is just the estimation performance of the Kalman-like estimator introduced earlier. To study the trade-off between estimation accuracy and energy consumption, we define the *total cost*

$$g_1^{\lambda}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \triangleq (1-\lambda)\,\mathrm{MSE}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) + \lambda \mathcal{E}\left(\mathbf{u}_{k-1}\right), \tag{8.32}$$

where $\lambda \in [0,1]$. Based on these definitions, we can define the following optimization problem:

$$\min_{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{L-1}} \mathbb{E}\left\{\sum_{k=1}^{L} g_1^{\lambda}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right)\right\}. \tag{8.33}$$

Here, $L < \infty$ is the horizon length. As before, our goal is to find an admissible sensor selection strategy, that is, a sequence of control inputs $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{L-1}$ the optimization problem in (8.33). This corresponds to a *finite-horizon, partially observable stochastic control problem*. However, in this case, the cost function depends on the observations sequence, instead of the state sequence, complicating the derivation of the optimal strategy.

8.6.3 Partial Observability and Sufficient Statistics

As already discussed in Section 8.4.3, imperfect state information problems are characterized by the expanding dimension of the information vector I_{k-1} during decision making at each time step. To address this issue, we seek a sufficient statistic for control purposes [27]. In contrast to standard POMDP models, where the sufficient statistic is the belief state, in our case, we can prove by induction that a suitable sufficient statistic is the probability distribution of $\mathbf{x}_k$ given I_{k-1}, that is,

$$\hat{\mathbf{p}}_k^- = \left[\hat{p}_k^{-,1}, \hat{p}_k^{-,2}, \ldots, \hat{p}_k^{-,n} \right]^{\mathrm{T}} \in \mathcal{P}. \tag{8.34}$$

This corresponds to the one-step predicted estimate of the system state, and we will refer to it hereafter as the *predicted belief state*. In one time step, the predicted belief state's evolution is described by Bayes' rule as follows:

$$\hat{\mathbf{p}}_{k+1}^- = \frac{\mathbf{Pr}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \hat{\mathbf{p}}_k^-}{\mathbf{1}_n^{\mathrm{T}} \mathbf{r}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \hat{\mathbf{p}}_k^-}. \tag{8.35}$$

Here, $\mathbf{r}\left(\mathbf{y}_k, \mathbf{u}_{k-1}\right) \triangleq \mathrm{diag}\left(f\left(\mathbf{y}_k \mid \mathbf{e}_1, \mathbf{u}_{k-1}\right), \ldots, f\left(\mathbf{y}_k \mid \mathbf{e}_n, \mathbf{u}_{k-1}\right) \right)$ is the $n \times n$ diagonal matrix of measurement vector probability density functions.

8.6.4 Smoothing Estimators

In Section 8.6.1, we presented the recursive expressions of an approximate MMSE estimator of the individual's physical activity. Herein, we obtain better state estimates by exploiting the availability of both past and future measurement vectors and control inputs. Theorem 8.4 gives the recursive expressions related to an approximate MMSE smoother for estimating the physical activity of the individual at each time step.

Theorem 8.4 Zois *et al.* [32]
The R-stage, approximate smoothed estimator of $\mathbf{x}_k$*, denoted by* $\hat{\mathbf{p}}_k^R$ *with* $R \geq k+1, k \geq 0$*, is given by the expression*

$$\hat{\mathbf{p}}_k^R = \hat{\mathbf{p}}_k + \sum_{s=k+1}^{R} \mathbf{C}_s \left(\mathbf{y}_s - \mathbf{y}_s^- \right) \tag{8.36}$$

with

$$\mathbf{C}_s = \left(\Theta_{k,s} - \hat{\mathbf{p}}_k^{s-1} \hat{\mathbf{p}}_s^{-,\mathrm{T}} \right) \mathcal{M}^{\mathrm{T}}\left(\mathbf{u}_{s-1}\right) \left(\mathcal{M}\left(\mathbf{u}_{s-1}\right) \bar{\sum_s} \mathcal{M}^{\mathrm{T}}\left(\mathbf{u}_{s-1}\right) + \tilde{\mathbf{Q}}_s \right)^{-1}, \tag{8.37}$$

where

$$\Theta_{k,s} = \mathbb{E}\left\{ \mathbf{x}_k \mathbf{x}_{s-1}^{\mathrm{T}} \mid \mathcal{F}_{s-1} \right\} \mathbf{P}^{\mathrm{T}}, \tag{8.38}$$

$$\mathbb{E}\left\{\mathbf{x}_k\mathbf{x}_{s-1}^{\mathrm{T}} \mid \mathcal{F}_{s-1}\right\} = \frac{\Theta_{k,s-1}\mathbf{r}\left(\mathbf{y}_{s-1},\mathbf{u}_{s-2}\right)}{\mathbf{1}_n^{\mathrm{T}}\left(\Theta_{k,s-1}\mathbf{r}\left(\mathbf{y}_{s-1},\mathbf{u}_{s-2}\right)\right)\mathbf{1}_n}, \tag{8.39}$$

with $\mathbb{E}\left\{\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}} \mid I_0\right\} = \mathrm{diag}\left(\hat{\mathbf{p}}_0\right)$ *and* $\tilde{\mathbf{Q}}_s \sum_{i=1}^{n} \hat{p}_s^{-,i}\mathbf{Q}_{i,\mathbf{u}_{s-1}}$.

The smoother's MSE performance can be determined similarly to the Kalman filter's MSE performance and is given by the *conditional smoothing error covariance matrix*

$$\sum_k^R \triangleq \mathbb{E}\left\{\left(\mathbf{x}_k - \hat{\mathbf{p}}_k^R\right)\left(\mathbf{x}_k - \hat{\mathbf{p}}_k\right)^{R,\mathrm{T}} \mid \mathcal{F}_R\right\} = \mathrm{diag}\left(\hat{\mathbf{p}}_k^R\right) - \hat{\mathbf{p}}_k^R\hat{\mathbf{p}}_k^{R,\mathrm{T}}, \quad R \geq k+1,\ k \geq 0. \tag{8.40}$$

In addition, we observe that the smoother gain matrix $\mathbf{C}_s$ depends nonlinearly on the measurement vectors, yielding a nonlinear smoother. For the expressions of the *fixed-point*, *fixed-interval*, and *fixed-lag* smoothers and related information, the interested reader is referred to Ref. [32].

8.7 SENSOR SELECTION STRATEGIES FOR THE ALTERNATIVE FORMULATION

8.7.1 Optimal Sensor Selection

The optimal sensor selection strategy of the optimization problem (8.33) is solved via the DP algorithm of Theorem 8.5.

Theorem 8.5 Zois [33]
For $k = L-1,\ldots,1$, *the cost-to-go function* $\overline{J}_k\left(\hat{\mathbf{p}}_k^-\right)$ *is related to* $\overline{J}_{k+1}\left(\hat{\mathbf{p}}_{k+1}^-\right)$ *through the recursion*

$$\overline{J}_k\left(\hat{\mathbf{p}}_k^-\right) = \min_{\mathbf{u}_{k-1}\in\mathcal{U}}\left[\ell\left(\hat{\mathbf{p}}_k^-,\mathbf{u}_{k-1}\right) + \int \mathbf{1}_n^{\mathrm{T}}\mathbf{r}\left(\mathbf{y},\mathbf{u}_{k-1}\right)\hat{\mathbf{p}}_k^-\overline{J}_{k+1}\left(\frac{\mathbf{Pr}\left(\mathbf{y},\mathbf{u}_{k-1}\right)\hat{\mathbf{p}}_k^-}{\mathbf{1}_n^{\mathrm{T}}\mathbf{r}\left(\mathbf{y},\mathbf{u}_{k-1}\right)\hat{\mathbf{p}}_k^-}\right)d\mathbf{y}\right], \tag{8.41}$$

where

$$\ell\left(\hat{\mathbf{p}}_k^-,\mathbf{u}_{k-1}\right) = \left(1-\lambda\right)\mathrm{tr}\left(\left(\mathbf{I} - \mathbf{G}_k\mathcal{M}\left(\mathbf{u}_{k-1}\right)\right)\sum\nolimits_k^-\right) + \lambda\mathcal{E}\left(\mathbf{u}_{k-1}\right). \tag{8.42}$$

with $\mathbf{G}_k$ *and* $\sum_k^-$ *depending on the predicted belief state* $\hat{\mathbf{p}}_k^-$. *The cost-to-go function for* $k = L$ *is given by*

$$\overline{J}_L\left(\hat{\mathbf{p}}_L^-\right) = \min_{\mathbf{u}_{L-1}\in\mathcal{U}}\left[\ell\left(\hat{\mathbf{p}}_L^-,\mathbf{u}_{L-1}\right)\right]. \tag{8.43}$$

The DP equations in (8.41) and (8.43) face the same problems, that is, uncountably infinite predicted belief space, exponentially large control space, as the ones in

Section 8.5.1. Additionally, we need to perform an intensive N-dimensional integration for a measurement vector of length N for each control input and each time step. However, what makes our problem fundamentally different from standard POMDP models and prevents the direct application of preexisting methods is the fact that cost-to-go functions (8.41) and (8.43) are nonlinear functions of the predicted belief state. For sufficient conditions regarding the structure of these functions that yield optimal efficient implementations, the reader is referred to Ref. [33]. In Sections 8.7.2 and 8.7.3, we discuss low-complexity, suboptimal strategies that can be implemented and run in real time.

8.7.2 Myopic Strategy

We start with the DP expressions in (8.41) and (8.43) and design the following myopic sensor selection strategy,

$$\mathbf{u}_k^{\text{myopic}} = \arg\min \ell\left(\hat{\mathbf{p}}_{k+1}^{-}, \mathbf{u}_k\right), \tag{8.44}$$

which chooses the control input with the minimum immediate cost, ignoring its long-term impact. This approach avoids the burden of the multidimensional integration in (8.41), but still faces the control space and nonlinearity issues. Still, in certain cases [33], it is optimal to divide the predicted belief space into distinct regions and execute the optimal sensor selection strategy as a *threshold test*, that is, examine in which region the predicted belief state falls into and use the associated control input.

8.7.3 Cost-Efficient Weiss–Weinstein Lower Bound Strategy

The nonlinear structure of (8.41) and (8.43) challenges both the theoretical analysis and implementation of the optimal sensing strategy. In this section, we devise an alternative sensor selection strategy that replaces the estimation accuracy metric, which yields the nonlinearity, with a lower bound on MSE that is easier to handle.

Our proposed algorithm is in fact a myopic sensor selection strategy that optimizes the trade-off between the sequential Weiss–Weinstein lower bound (WWLB) [33, 34] and the energy consumption metric at each time step,

$$\mathbf{u}_k^{\text{CE-WWLB}} = \arg\min\left[\left(1-\lambda\right) v\left(\mathbf{u}_k\right) + \lambda c\left(\mathbf{u}_k\right)\right], \tag{8.45}$$

where $v\left(\mathbf{u}_k\right) \triangleq \max_{\mathbf{h}_{k+1}}\left[J_{k+1}^{-1}\left(\mathbf{h}_{k+1}, \mathbf{u}_k\right)\right]$ is the tightest WWLB for a particular control input and $\mathbf{h}_{k+1}$ denotes the test point (see discussion later) selected at time step $k+1$. We refer to this algorithm as cost-efficient WWLB (CE-WWLB) strategy.

For our system model, the sequential WWLB at time step k is given by $\mathbb{E}\left\{\left(\mathbf{x}_k - \hat{\mathbf{p}}_k\right)\left(\mathbf{x}_k - \hat{\mathbf{p}}_k\right)^{\mathrm{T}}\right\} \geq J_k^{-1}$ and can be updated as follows

$$A_{k+1} = B_{k,k}^{k+1} - B_{k,k-1}^{k} A_k^{-1} B_{k,k-1}^{k}, \quad \forall k = 0,1,\ldots, \tag{8.46}$$

$$J_{k+1} = B_{k+1,k+1}^{k+1} - B_{k+1,k}^{k+1} A_{k+1}^{-1} B_{k+1,k}^{k+1}, \quad \forall k = 0,1,\ldots, \tag{8.47}$$

with initial conditions $A_0^{-1} \triangleq 0, B_{0,-1}^0 \triangleq 0$ and $B_{k,k}^{k+1}, B_{k+1,k+1}^{k+1}, B_{k+1,k}^{k+1}$ given by Lemma 8.5.

Lemma 8.5 Zois [33]
Let $P(\mathbf{x}_0)$ be known a priori probability mass function related to the initial state $\mathbf{x}_0$. Then, the sequential WWLB at each time step k is determined by (8.46) *and* (8.47)*, where*

$$B_{k+1,k+1}^{k+1} = \frac{2\left(1-\exp\left(\eta_k\left(\mathbf{h}_{k+1},-\mathbf{h}_{k+1},\mathbf{u}_k\right)\right)\right)}{\exp\left(2\eta_k\left(h_{k+1},0,\mathbf{u}_k\right)\right)}, \tag{8.48}$$

$$\begin{aligned} B_{k+1,k}^{k+1} = {} & \frac{\exp\left(\zeta_k\left(\mathbf{h}_k,\mathbf{h}_{k+1},\mathbf{u}_{k-1},\mathbf{u}_k\right)\right)-\exp\left(\zeta_k\left(-\mathbf{h}_k,\mathbf{h}_{k+1},\mathbf{u}_{k-1},\mathbf{u}_k\right)\right)}{\exp\left(\eta_k\left(\mathbf{h}_{k+1},0,\mathbf{u}_k\right)+\xi_k\left(\mathbf{h}_k,0,\mathbf{u}_{k-1}\right)\right)} \\ & + \frac{\exp\left(\zeta_k\left(-\mathbf{h}_k,-\mathbf{h}_{k+1},\mathbf{u}_{k-1},\mathbf{u}_k\right)\right)-\exp\left(\zeta_k\left(\mathbf{h}_k,-\mathbf{h}_{k+1},\mathbf{u}_{k-1},\mathbf{u}_k\right)\right)}{\exp\left(\eta_k\left(\mathbf{h}_{k+1},0,\mathbf{u}_k\right)+\xi_k\left(\mathbf{h}_k,0,\mathbf{u}_{k-1}\right)\right)}, \end{aligned} \tag{8.49}$$

$$B_{k,k}^{k+1} = \frac{2\left(1-\exp\left(\xi_k\left(\mathbf{h}_k,-\mathbf{h}_k,\mathbf{u}_{k-1}\right)\right)\right)}{\exp\left(2\xi_k\left(\mathbf{h}_k,0,\mathbf{u}_{k-1}\right)\right)}, \tag{8.50}$$

with

$$\begin{aligned} \eta_k\left(\mathbf{h}_a,\mathbf{h}_b,\mathbf{u}\right) = {} & \ln\sum_{\mathbf{x}_k} P\left(\mathbf{x}_k\right)\sum_{\mathbf{x}_{k+1}}\sqrt{P\left(\mathbf{x}_{k+1}+\mathbf{h}_a \mid \mathbf{x}_k\right)}\sqrt{P\left(\mathbf{x}_{k+1}+\mathbf{h}_b \mid \mathbf{x}_k\right)} \\ & \times \rho\left(\mathbf{x}_{k+1}+\mathbf{h}_a,\mathbf{x}_{k+1}+\mathbf{h}_b,\mathbf{u}\right), \end{aligned} \tag{8.51}$$

$$\begin{aligned} & \xi_k\left(\mathbf{h}_a,\mathbf{h}_b,\mathbf{u}\right) \\ & = \ln\sum_{\mathbf{x}_{k-1}} P\left(\mathbf{x}_{k-1}\right)\sum_{\mathbf{x}_k}\sqrt{P\left(\mathbf{x}_k+\mathbf{h}_a \mid \mathbf{x}_{k-1}\right)}\sqrt{P\left(\mathbf{x}_k+\mathbf{h}_b \mid \mathbf{x}_{k-1}\right)}\sum_{\mathbf{x}_{k+1}}\sqrt{P\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k+\mathbf{h}_a\right)} \\ & \times\sqrt{P\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k+\mathbf{h}_b\right)}\rho\left(\mathbf{x}_k+\mathbf{h}_a,\mathbf{x}_k+\mathbf{h}_b,\mathbf{u}\right), \end{aligned} \tag{8.52}$$

$$\begin{aligned} & \zeta_k\left(\mathbf{h}_a,\mathbf{h}_b,\mathbf{u}_a,\mathbf{u}_b\right) \\ & = \ln\sum_{\mathbf{x}_{k-1}} P\left(\mathbf{x}_{k-1}\right)\sum_{\mathbf{x}_k}\sqrt{P\left(\mathbf{x}_k+\mathbf{h}_a \mid \mathbf{x}_{k-1}\right)P\left(\mathbf{x}_k \mid \mathbf{x}_{k-1}\right)}\sum_{\mathbf{x}_{k+1}}\sqrt{P\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k+\mathbf{h}_a\right)} \\ & \times\sqrt{P\left(\mathbf{x}_{k+1}+\mathbf{h}_b \mid \mathbf{x}_k\right)}\rho\left(\mathbf{x}_k+\mathbf{h}_a,\mathbf{x}_k,\mathbf{u}_a\right)\rho\left(\mathbf{x}_{k+1}+\mathbf{h}_b,\mathbf{x}_{k+1},\mathbf{u}_b\right), \end{aligned} \tag{8.53}$$

and the function $\rho(\cdot)$ corresponds to the Bhattacharyya coefficient in (8.5)*. Furthermore, the information submatrix J_0 at time step* 0 *is*

$$J_0 = \frac{2\left(1-\exp\left(\gamma\left(\mathbf{h}_0,-\mathbf{h}_0,\mathbf{u}_{-1}\right)\right)\right)}{\exp\left(2\gamma\left(\mathbf{h}_0,0,\mathbf{u}_{-1}\right)\right)} \tag{8.54}$$

with $\gamma\left(\mathbf{h}_a,\mathbf{h}_b\right) = \ln\sum_{\mathbf{x}_0}\sqrt{P\left(\mathbf{x}_0+\mathbf{h}_a\right)P\left(\mathbf{x}_0+\mathbf{h}_b\right)}\rho\left(\mathbf{x}_0+\mathbf{h}_a,\mathbf{x}_0+\mathbf{h}_b,\mathbf{u}\right)$.

WWLB adopts the usage of test points in an effort to overcome certain differentiability conditions, known as *regularity conditions*, which are indispensable to other well-known MSE bounds such as the Cramér–Rao and the Bhattacharyya lower bound [34]. This suggests that it can be used in the estimation of discrete parameters, as in our case. However, the special structure of the parameter space $\mathcal{X}$ indicates that test points should be state dependent to establish the validity and correctness of the related expressions. For instance, the test points for state $\mathbf{e}_2$ should be selected such that the resulting states are $\mathcal{X} - \{\mathbf{e}_2\}$.

The CE-WWLB strategy has various nice features. First, the WWLB metric $v(\mathbf{u}_k)$ can be expressed as a function of union-bound terms on the detection probability. As a result, we determine a connection between MSE performance and detection accuracy, which we considered earlier in Section 8.4.2. Second, we avoid both the evaluation of the multidimensional integration in (8.41) and the optimization over the predicted belief space in (8.41), (8.43), and (8.44) since $v(\mathbf{u}_k)$ does not depend on the predicted belief state. Last, the latter independence also suggests that the CE-WWLB strategy can be determined off-line and implemented as a look-up table in real time.

8.8 EXPERIMENTS

In this section, we illustrate the performance of the sensor selection strategies presented in Sections 8.5 and 8.7. Our simulations are based on the experimental data collected by the KNOWME network and the energy costs experimentally determined during its operation.

8.8.1 Simulations Framework

Data collection via the KNOWME network system was conducted in the lab and consisted of three to four sessions, where 12 individuals were asked to perform eight specific physical activities. For a thorough description of the data collection process, protocols, and individual characteristics, we refer the reader to Ref. [26]. Herein, we present results for two individuals only. Our aim is to estimate the time-evolving physical activity of each individual by distinguishing between four activities: *Sit*, *Stand*, *Run,* and *Walk*. We use for both individuals the same Markov chain, which is shown in Figure 8.9. From the biometric signals of each of the WBAN sensors, we use only one extracted feature—(i) the ACC mean (average acceleration signal means in each axis for a data window) from the mobile phone's ACC, (ii) the ACC variance (average acceleration signal variances in each axis for a data window) from the external ACC, and (iii) the ECG period (ECG waveform interpeak period) from the external ECG. We underscore that these features are part of the optimal feature set established by the feature selection methods in Ref. [26]. Furthermore, we currently do not use GPS and OXI readings as they have been proven to be less effective for physical activity detection compared to other sensor measurements. The feature distributions for the four physical activities of the two individuals are illustrated in Figure 8.10.

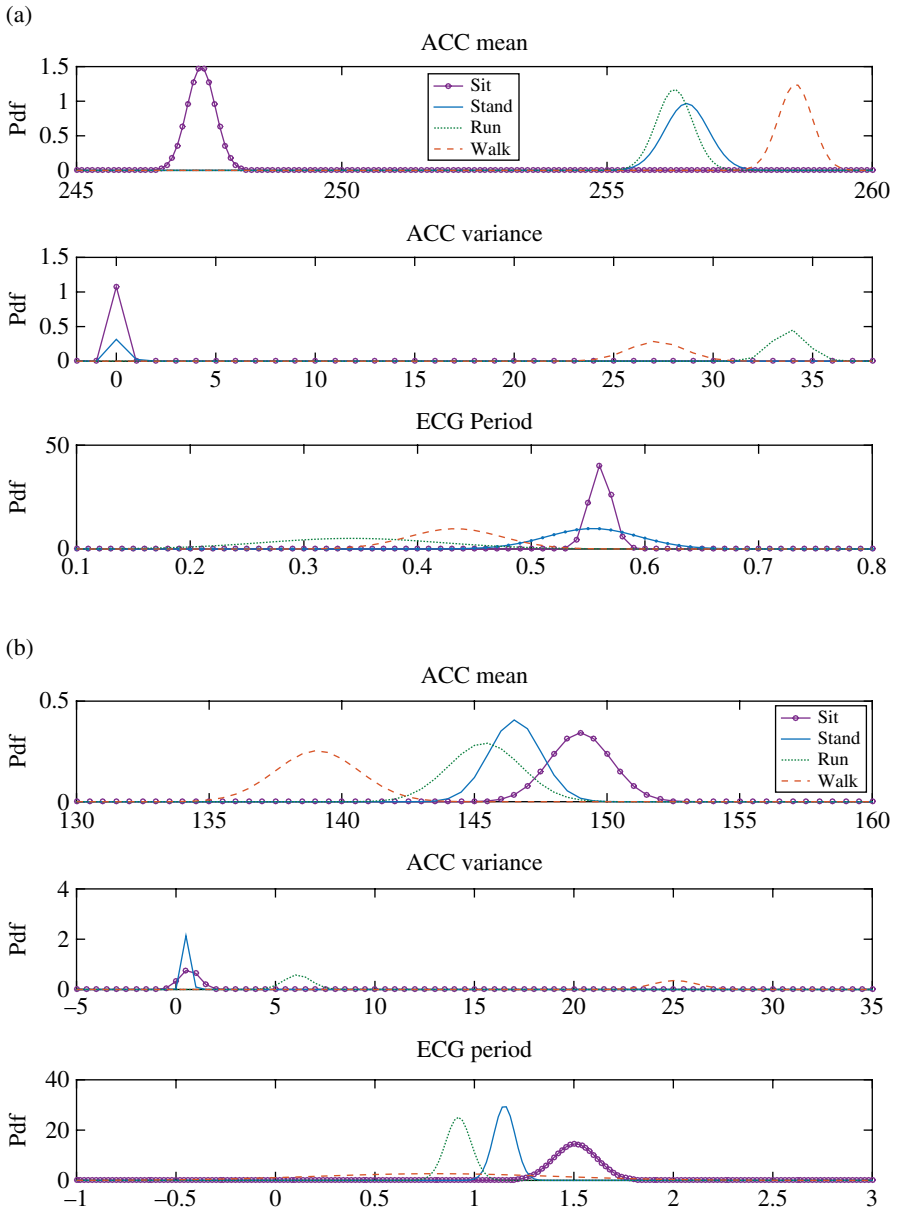


FIGURE 8.10 Feature distributions for the four physical activities of two different individuals: (a) individual 1 and (b) individual 2.

Unless otherwise stated, all results were acquired for total number of samples $N = 12$, horizon length $L = 5$ and were averaged over 10^6 Monte Carlo runs. The horizon length was empirically determined by considering the impact of the sensor selection strategy on the estimation of future physical activities. Finally, based on the study presented in Section 8.3, we selected $\delta = [\delta_{\text{ACC1}}, \delta_{\text{ACC2}}, \delta_{\text{ECG}}]^{\text{T}} = [0.585, 0.776, 1]^{\text{T}}$, where ACC 1 denotes the mobile phone's ACC and ACC 2 the external ACC. In the sequel, as a baseline, we use an equal allocation (EA) strategy for $N = 3, 6, 9, 12$, which receives equal number of samples from each sensor without considering the individual's physical activity. We compute the following metrics:

1. *Average detection performance* (ADP):

$$\text{ADP} \triangleq \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\{\mathbf{x}_k = \hat{\mathbf{x}}_k\}}; \tag{8.55}$$

2. *Average MSE performance* (AMSE):

$$\text{AMSE} \triangleq \frac{1}{K} \sum_{k=1}^{K} \text{tr}\left(\sum_{k} \right); \tag{8.56}$$

3. *Average energy cost* (AEC):

$$\text{AEC} \triangleq \frac{1}{K} \sum_{k=1}^{K} \mathbf{u}_k^{\text{T}} \delta, \tag{8.57}$$

where K represents the number of Monte Carlo runs.

8.8.2 Numerical Results

As verified by Figure 8.10, choosing to receive measurements from a few sensors enables us to distinguish between a subset of the physical activities. Our simulation results confirm this observation and, in fact, the selected number of control inputs is much less than the total number of available ones. The appropriate number of samples at each step is a function of the desired level of accuracy and energy consumption. Based on these observations, we may be able to design suitable heuristics that determine the appropriate subset of control inputs for each individual.

Figure 8.11a illustrates the AEC–ADP trade-off of DP and MIC-T3S for the two individuals and the performance of EA. We observe that DP and MIC-T3S exhibit very similar performance, which varies with energy consumption, that is, detection accuracy ameliorates as we spend more energy. In fact, we observe energy gains ranging from 29 to 74% for the same detection performance as EA. Thus, sensor selection can lead to significant energy savings versus EA. Of course, DP and MIC-T3S achieve much higher maximum detection performance than EA inasmuch as they consider the individual's current physical activity. On the other hand, the former strategies can achieve 100% energy gains by sacrificing detection accuracy. Figure 8.11b demonstrates the AEC–ADP trade-off curves of E^{2}MBADP and

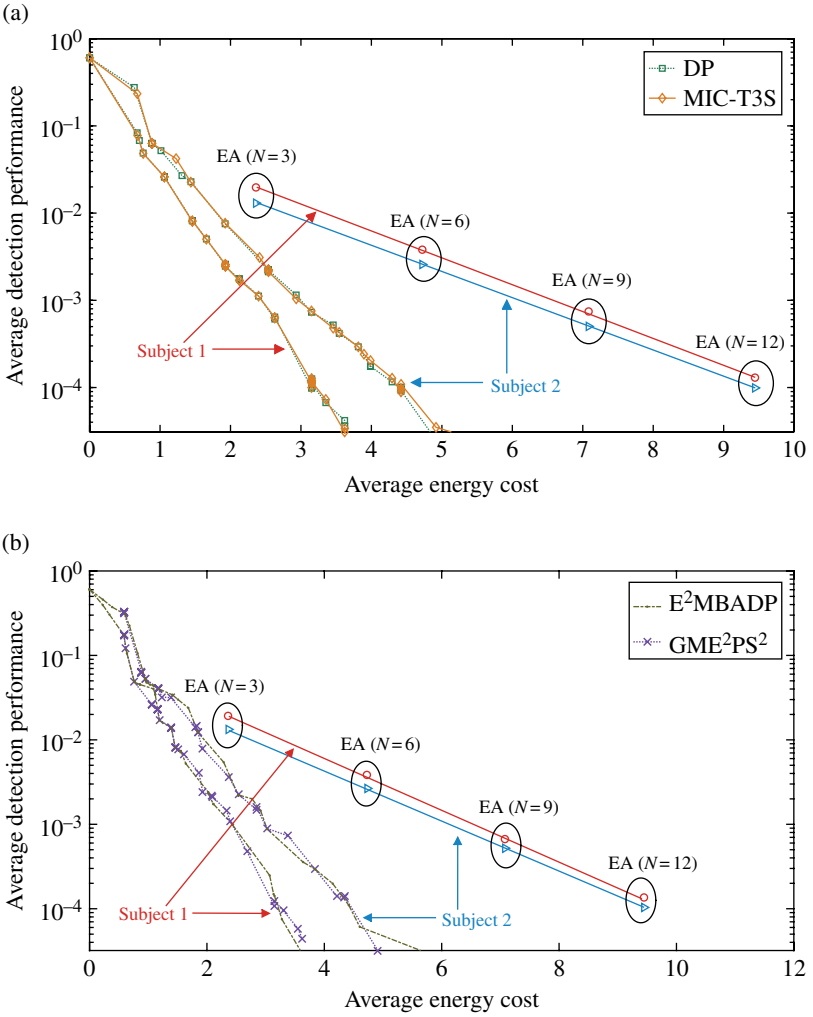


FIGURE 8.11 Trade-off between ADP and AEC for $N = 12$ samples and two individuals. (a) DP and MIC-T3S strategies and (b) E²MBADP and GME²PS² strategies.

GME²PS² for the two individuals. In general, we notice similar trends as in the case of DP and MIC-T3S. Energy savings, however, are slightly different ranging from 22 to 66 % (68 %) for E²MBADP (GME²PS²) for the same detection performance as EA. In addition, in certain cases, GME²PS² achieves higher detection performance with higher energy consumption, as opposed to E²MBADP, which is due to their distinct performance metrics. Finally, due to user variability, all strategies exhibit different energy savings, which increase with N, for the two individuals.

Figure 8.12 shows the average number of samples selected by MIC-T3S from each sensor with respect to λ for the two individuals. We notice that due to used

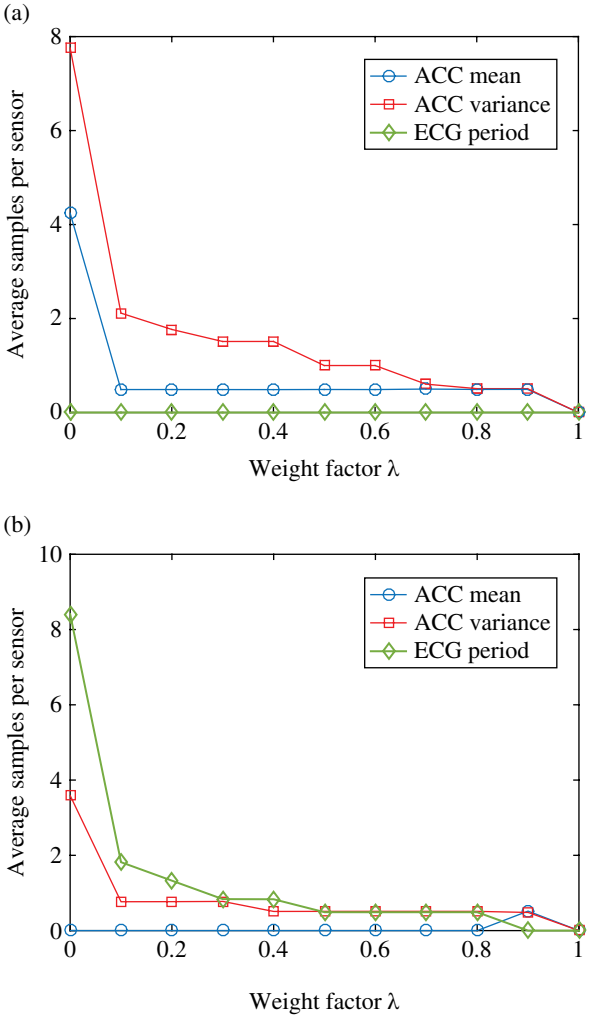


FIGURE 8.12 Average number of samples selected by MIC-T3S from each sensor for $N = 12$ samples and varying values of λ for the two individuals: (a) individual 1 and (b) individual 2.

variability, we have significantly different allocation of samples. For instance, no ECG samples are used for individual 1 versus individual 2. Figure 8.13 presents the average number of samples selected by E^{2}MBADP and GME^{2}PS^2 from each sensor with respect to $\tau \in [0.95,1]$ for the two individuals. For $\tau < 0.95$, GME^{2}PS^2 always selects at least one sample, contrary to E^{2}MBADP, which sometimes decides to receive no samples. In any case, both strategies receive most samples from ACC mean for low values of τ since it is the most energy-efficient WBAN sensor.

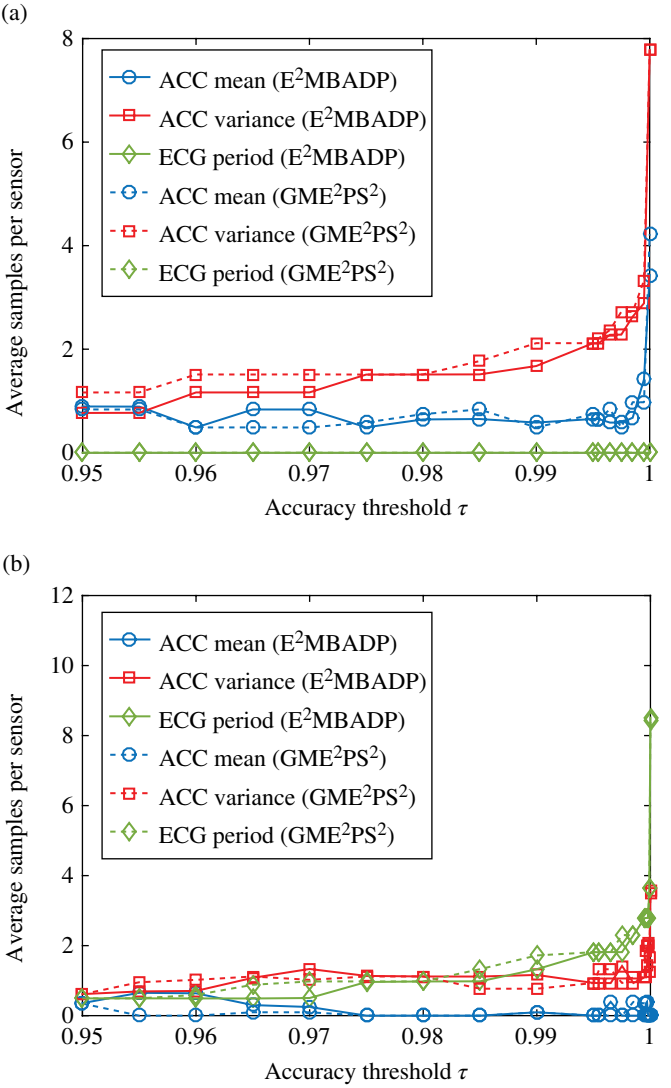


FIGURE 8.13 Average number of samples selected by E²MBADP and GME²PS² from each sensor for $N = 12$ samples and varying values of τ for the two individuals: (a) subject 1 and (b) subject 2.

For larger values of τ, samples are selected from the rest of the sensors, and any differences can be attributed to the different detection metrics employed by the strategies.

Figure 8.14 shows the average number of samples for each physical activity selected from each WBAN sensor by the sensor selection strategies in Section 8.5. Results are shown for same detection performance as EA. We notice that sample selection is a function of the physical activity and the individual. Furthermore, we see that less than

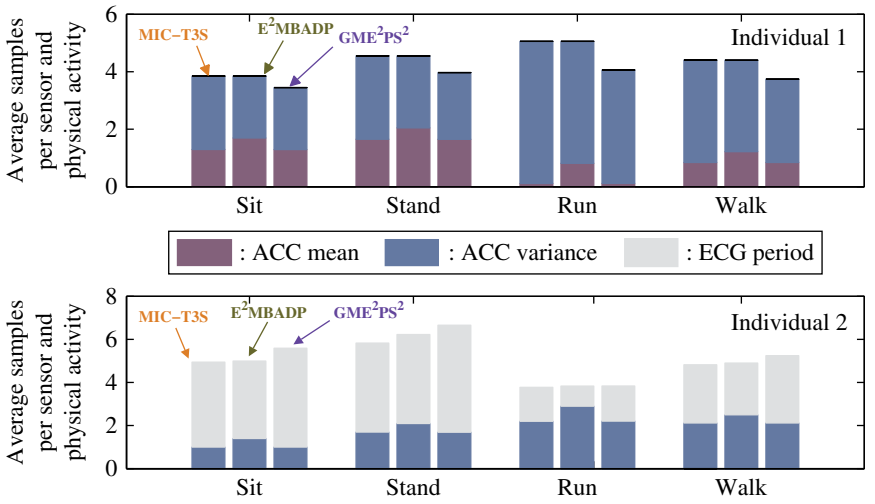


FIGURE 8.14 Average number of samples selected from each sensor for each physical activity by MIC-T3S, E²MBADP, and GME²PS².

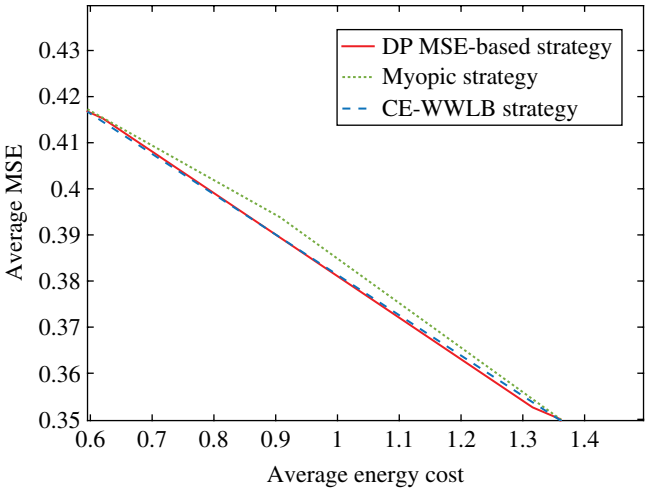


FIGURE 8.15 Trade-off between AMSE and AEC for DP, myopic and CE-WWLB strategies for $N = 2$ samples.

N samples are selected in total (i.e., we have a 69 % reduction on sample usage) since more informative samples are used. This gives rise to important energy savings, which, however, depend on the specific individual and the physical activity performed. Finally, we notice that MIC-T3S and E²MBADP employ similar sample selection strategies versus GME²PS², which is attributed to the greedy nature of the latter.

Figure 8.15 illustrates the AEC–AMSE trade-off of DP in (8.41)–(8.43), myopic algorithm in (8.44), and CE-WWLB in (8.45) for individual 1. The total number N of

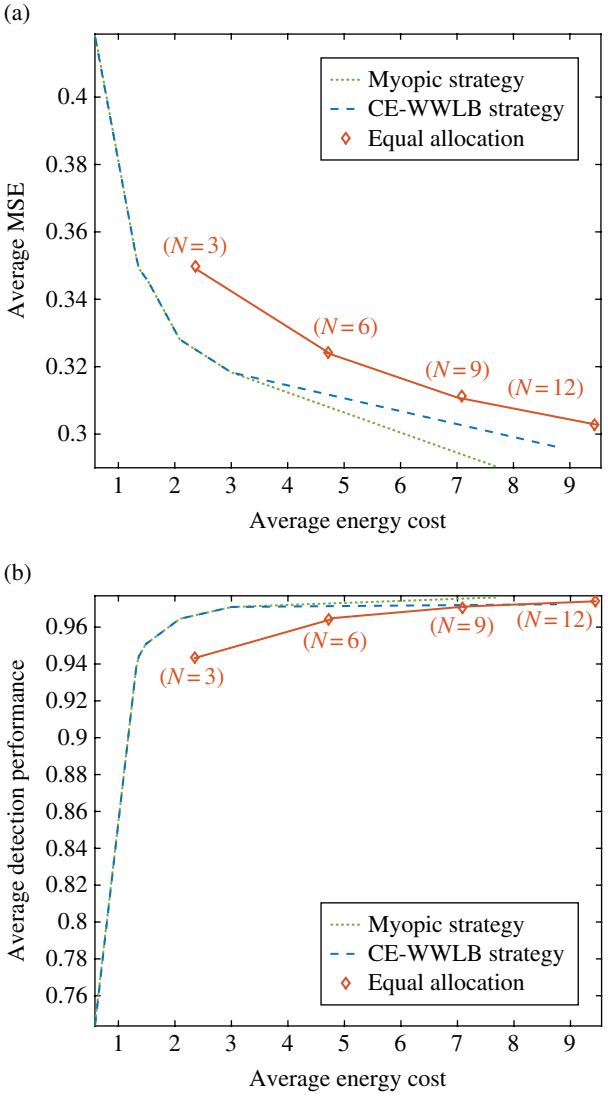


FIGURE 8.16 Trade-off curves for myopic and CE-WWLB strategies. (a) AMSE versus AEC and (b) ADP versus AEC.

samples was set to two, since for larger values of N, DP involves significant computation time. We observe that the suboptimal strategies' performance is very close to the optimal DP solution. CE-WWLB actually coincides with DP, while for the myopic algorithm we have a slight performance loss. We believe that CE-WWLB's outstanding performance is due to the connection between MSE and detection performance captured by the WWLB metric. Next, Figure 8.16 demonstrates the trade-off of the myopic and CE-WWLB algorithms and the performance of EA for

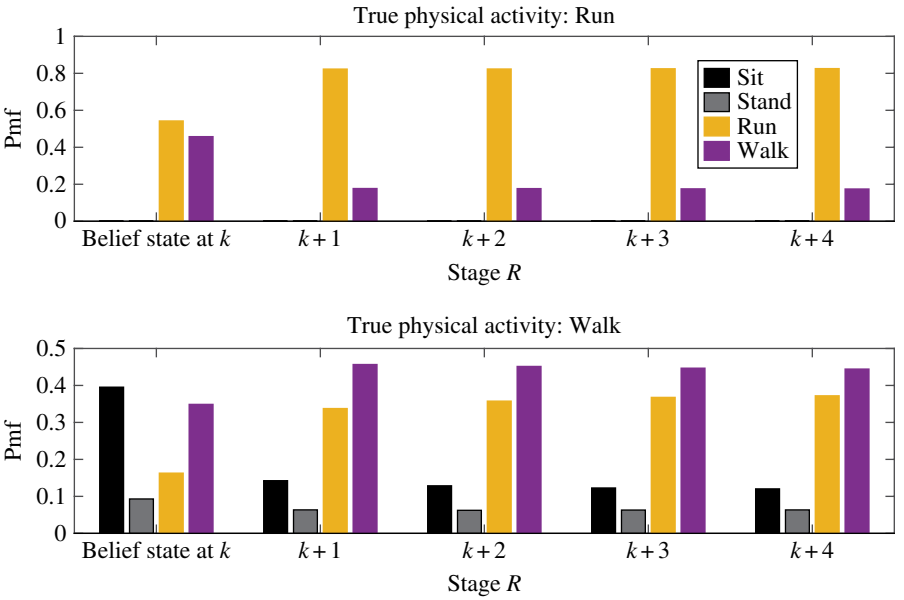


FIGURE 8.17 Effect of smoothing on the belief state for increasing values of R.

individual 1. More specifically, Figures 8.16a and b present the AEC–AMSE and AEC–ADP trade-offs, respectively. We notice that both algorithms exhibit similar detection/estimation performance as EA but with lower energy consumption. In general, we can attain better performance by consuming more energy, and the two algorithms follow roughly the same trend. However, when maximum detection/estimation performance is desired, the myopic algorithm performs better than CE-WWLB by spending more energy. This is a direct outcome of the latter algorithm-making decisions without using the predicted belief state values. Therefore, we observe only 7 % energy savings for CE-WWLB versus 60 % for the myopic algorithm of the same performance as EA.

At this point, we briefly discuss about the effect of smoothing on the improvement of detection performance. In particular, depending on the individual's statistics and Markov chain of physical activities, we may achieve up to 2 % improvement, which of course saturates as the stage R grows. Figure 8.17 shows the effect of stage R on the belief state estimate. In particular, we see that future information can indeed enhance or overturn our belief regarding the individual's true physical activity. Finally, we notice that MSE does improve significantly with smoothing contrary to detection performance.

Last but not least, Figure 8.18a and b illustrate the average number of samples selected from each sensor by the myopic and CE-WWLB strategies, respectively, for same detection performance as EA. Results are shown for individual 1. We notice that both algorithms do not request any samples from ECG period, as anticipated

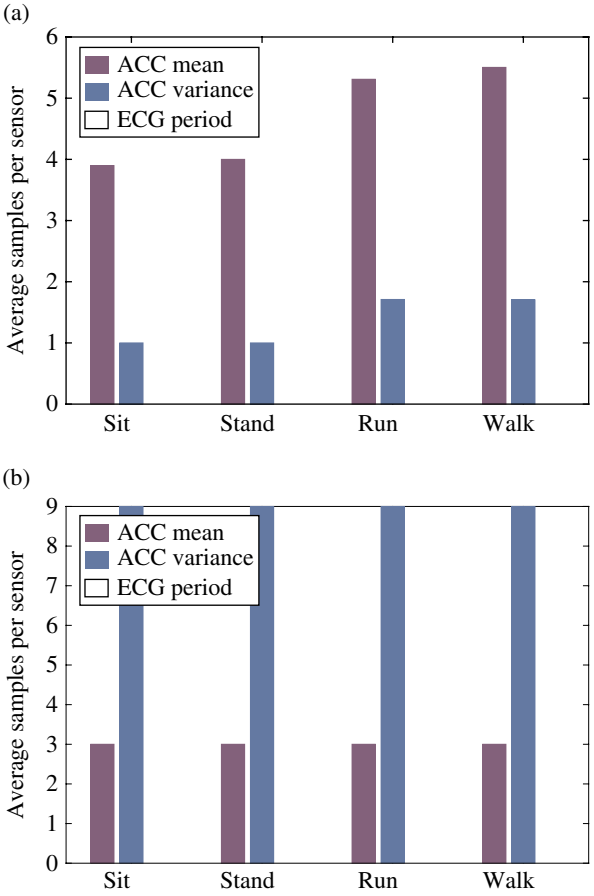


FIGURE 8.18 Average number of samples selected from each sensor for each physical activity by myopic and CE-WWLB. (a) Myopic strategy and (b) CE-WWLB strategy.

from Figure 8.10a. Instead, they choose to receive different number of samples from the other two WBAN sensors. Specifically, the myopic strategy selects more samples from ACC mean and in general, less than N. On the other hand, the CE-WWLB strategy always selects the same number of samples from all sensors (mostly from ACC mean) irrespective of the individual's physical activity inasmuch as it ignores belief state information.

In summary, we observe that the design of smart sensor selection strategies, such as the ones presented in Sections 8.5 and 8.7, can yield substantial energy savings and significantly prolong the lifetime of WBANs without sacrificing detection/estimation performance. In addition, the use of such strategies in real WBANs is feasible because of their low computational complexity. In any case, it is imperative to perform personalized training during the design process since use variability plays a

crucial role on the achieved energy gains. For a discussion on the effect of the Markov chain transition matrix on the performance of the proposed strategies and methods for detecting ephemeral states, the reader is referred to Ref. [33].

8.9 RELATED WORK

Energy-efficient operation of wireless sensor networks (WSNs) is an extensively studied problem in the literature [35]. Since sensors usually correspond to battery-powered devices, prolonging network lifetime can be achieved by reducing power consumption in the node level. Recently, there has been a deluge of energy-saving frameworks and approaches, for example, ad hoc algorithms, data-driven schemes, and duty-cycling/sensor selection methods, for a variety of sensor network applications.

Among the various approaches, sensor selection, which refers to the construction of a schedule under which sensor nodes interchange between active and sleep modes depending on the monitored process and/or network conditions, is a particularly attractive approach. Prior art has proposed several algorithmic [36–43] and mathematical sleep-scheduling frameworks [44–52] for various applications. Contrary to our problem, where the focus is the optimization of the energy resources in the fusion center, in all of the earlier cases (apart from Refs. [48, 49]), the objective is to optimize energy consumption in the sensor node level. In typical hierarchical algorithms [38], high-accuracy sensors are activated only when a critical event happens, while energy-efficient sensors are continuously monitoring the process of interest. In node-clustering methods [37, 39], sensor nodes compose groups, where a representative node is selected periodically by a coordinator or the rest of the nodes and remains awake to manage communication. In contrast, we choose sensors depending on their energy consumption and detection performance to estimate the individual's time-evolving physical activity, while there are situations where no sensors are selected at all.

In hand-held devices, various methods have been proposed to improve energy consumption. Viredaz *et al.* [53] proposed the detection of idle time periods in a mobile device and the use of voltage frequency scaling to reduce power consumption. Shih *et al.* [54] showed that using wake-on-wireless, a PDA can be put into sleep mode and woken up only on an incoming call or when the user is actively using the device. Turdecken [55] demonstrated how to use hierarchical power management to reduce energy consumption. In this context, a low-power mote is attached to a mobile node and used to continuously monitor incoming packets. The mote wakes up the mobile device when an incoming packet is detected. The notion of hierarchical energy management is exploited at a much finer granularity in EEMSS [12]. Namely, EEMSS categorizes all the sensors on a mobile phone into a hierarchy and then activates low-energy sensors that will in turn decide when to activate higher-energy intensive sensors.

In WBANs, energy-efficient physical activity detection has been explored through the use of intelligent feature selection methods and Bayesian statistics [36, 40].

In particular, sensors are selected based on their informational capabilities in a step-by-step manner to accurately infer the underlying, nonevolving human state. On the other hand, our approach builds upon a robust optimization framework, which ensures optimal behavior by evaluating both accuracy and energy requirements to detect the individual's time-evolving physical activity. In Ref. [56], an energy-efficient computing model is investigated for a WBAN employing Bluetooth or Zigbee to communicate with a central device. Energy-efficient context monitoring to detect the physical activity changes is addressed in Refs. [15, 57] via provisioning of application interfaces for sensor management. In Ref. [58], a sampling scheme that minimizes energy consumption is proposed based on the redistribution of unused time. In contrast, our work offers a new approach in that we achieve energy-efficient physical activity recognition every time step-by-step, optimizing the allocation strategy of the mobile phone via a principled optimization approach that enables us to consider the trade-off between informativeness and energy usage of each WBAN node.

In recent years, stochastic control approaches such as Markov decision processes (MDPs) have been exploited to choose sensor subsets [44] and transmission rates or modes [47–49]. For instance, in Refs. [48, 49], a constrained MDP formulation was introduced to determine the optimal sensor sampling policy asssuming missing observations and a constrained energy budget. The goal was to accurately estimate a Markovian/semi-Markovian time-evolving user state process. In contrast to the mentioned schemes, we are not aware of the true system state, but this is what we would like the WBAN to estimate. Energy efficiency in WSNs and WBANs has also been explored through POMDPs [45, 46, 50–52]. In our work, we adopt a system and a cost model that adheres to the properties of real WBAN systems versus [45, 46, 50, 51], where simplified system and cost models are considered based on which various approximate algorithms are designed. In fact, we acknowledge the sensor nodes' individual characteristics and suggest algorithms for realistic scenarios. The problem of energy-efficient classification in WBANs is considered in Ref. [52] and shares some similarities with our work. Even so, we do not optimize the energy consumption of the sensors due to data transmission, but the listening strategy of the mobile phone. On top of that, we work with heterogeneous sensors, consider various transmission rates contrary to activate/deactivate all sensors approach addressed in Ref. [52], and develop closed-form formulae for the detection accuracy and energy cost functions. Finally, we develop original approximation algorithms (versus using already existing ones) exploiting properties of our cost functions and heuristics that match the unique characteristics of our problem.

Recent studies have also performed energy measurement of wireless data transmission using various network interfaces [59]. It is shown that data transmission energy varies widely from one location to another and may also vary at the same location depending on the time. In Ref. [60], the authors introduce Wiffler, which is designed for optimizing data throughput in a vehicular network. Wiffler can change between Wi-Fi and 3G depending on network conditions, and it uses historical data to predict future available Wi-Fi access points. Since Wiffler is geared toward vehicular network, its focus is more on bandwidth, whereas a WBAN is sensitive to battery

consumption. BreadCrumbs [61] tracks user's movement to generate connectivity forecasts. Ra *et al.* [62] focused on dynamically selecting between various wireless radios. They introduced SALSA that uses a Lyapunov optimization framework to automatically decide when to send/defer data transmission to wait for better channel availability so as to optimize the overall energy-delay trade-offs. Again, many of these approaches use the prediction of Wi-Fi access points or user movement to optimize energy of at least one component that is used in KNOWME. We can carefully exploit these prediction models within the KNOWME context and take advantage of them to further reduce energy consumption.

While there are disparate sources of some of the energy consumption information, none of the previous studies have done a systematic and comprehensive analysis of the energy consumption of a WBAN starting with the initial system design choices to the data transmission. We believe that our research also sheds new light on issues such as programmability and energy efficiency, energy compression costs of sensor data, and the energy cost of storing data locally on mobile phones. Given the energy consumption uncertainty in WBANs, it is necessary to use a profile-based dynamic adaptation to minimize the energy consumption across all WBAN layers.

Most prior work has also adopted somewhat restrictive assumptions. First, "perfect sensing of state" (i.e., if data transmission is successful, the unknown process is fully observable [47–49, 51]) is typically assumed. Furthermore, sensor capabilities, usage costs, or both are the same for all sensors [37, 39, 44, 45, 47–52]. Last but most importantly, energy optimization happens in the sensor nodes, not the fusion center level [36–40, 44–47, 50–52]. Our problem, on the other hand, has the following distinct characteristics, which make imperative the design of new mathematical formulations and algorithms:

1. Sensors are heterogeneous in detection/estimation capabilities and energy cost (reception cost of one sample from each sensor[1]).
2. The sensors' sensing and communication modes introduce errors, and thus the time-evolving physical activity is observed through noisy measurements.
3. Only a limited number of heterogeneous sensors are used to ensure compliance in wear.
4. The mobile phone constitutes the energy bottleneck, not the sensors.

8.10 CONCLUSION

WBANs constitute the cornerstone of the e-health ecosystem and address various health-related problems. Thus, optimizing their operation is the necessary step to ensure viability of the myriads of applications. In this chapter, we argued that energy efficiency can significantly prolong the lifetime of WBANs. In particular, we studied the problem of energy-efficient physical activity detection and identified the effect of

[1] The data processing cost is outside of the focus of the current work.

various design choices on energy consumption and their impact. In addition, we devised a powerful formulation that addresses the challenge of optimizing energy consumption due to sensor data collection and captures the problem's key characteristics. This enabled us to develop appropriate theoretical results as well as optimal and near-optimal, low-complexity algorithms. Our extensive simulations suggest that energy efficiency in WBANs is a realistic goal that can be achieved without compromising the objectives of related applications.

ACKNOWLEDGMENTS

This work has been funded in part by the following grants and organizations: ONR N00014-09-1-0700, NSF CNS-0832186, NSF CCF-0917343, NSF CCF-1117896, NSF CNS-1213128, NSF CCF-1410009, CPS-1446901, AFOSR FA9550-12-1-0215, DOT CA-26-7084-00, the National Center on Minority Health and Health Disparities (NCMHD) (supplement to P60 MD002254), Nokia, and Qualcomm.

REFERENCES

[1] V. Della Mea, "What is e–Health (2): The death of telemedicine?" Journal of Medical Internet Research, 3(2), e22, 2001.

[2] D.-S. Zois, M. Levorato, and U. Mitra, "Energy-efficient, heterogeneous sensor selection for physical activity detection in wireless body area networks," IEEE Transactions on Signal Processing, 61(7), 1581–1594, 2013.

[3] E. Jovanov, A. Milenkovic, C. Otto, and P. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," Journal of NeuroEngineering and Rehabilitation, 2(1), 6, 2005.

[4] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," Proceedings of the IEEE, 101(5), 1203–1233, 2013.

[5] A. K. Bourke, J. V. O'Brien, and G. M. Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm," Gait & Posture, 26(2), 194–199, 2007.

[6] C. Di Natale, A. Macagnano, E. Martinelli, R. Paolesse, G. D'Arcangelo, C. Roscioni, A. Finazzi Agro, and A. D'Amico, "Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors," Elsevier: Biosensors and Bioelectronics, 18(10), 1209–1218, 2003.

[7] D. Spruijt-Metz, D. Berrigan, L. A. Kelly, R. McConnell, D. Dueker, G. Lindsey, A. Atienza, S. Michel, M. L. Irwin, J. Wolch, M. Jerrett, Z. Tatalovich, and S. Redline. Measures of physical activity and exercise. In: Handbook of Assessment Methods for Obesity and Eating Behaviors, Related Problems and Weight: Measures, Theory and Research, A. D. Allison (ed.), 2nd ed. Thousand Oaks, CA: Sage Publications Inc., 2008, pp. 187–254.

[8] B. Caballero, "The global epidemic of obesity: An overview," Epidemiologic Reviews, 29(1), 1–5, 2007.

[9] C. Ogden, M. Carroll, and K. Flegal, "High body mass index for age among us children and adolescents," Journal of the American Medical Association, 299(20), 2401–2405, 2008.

[10] U. Mitra, B. A. Emken, S. Lee, M. Li, V. Rozgic, G. Thatte, H. Vathsangam, D. Zois, M. Annavaram, S. Narayanan, M. Levorato, D. Spruijt-Metz, and G. Sukhatme, "KNOWME: A case study in wireless body area sensor network design," IEEE Communications Magazine, 50(5), 116–125, 2012.

[11] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. Leung, "Body area networks: A survey," Mobile Networks and Applications, 16, 171–193, 2011.

[12] Y. Wang, J. Lin, M. Annavaram, Q. A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh, "A framework of energy efficient mobile sensing for automatic user state recognition," in *MobiSys'09: Proceedings of the Seventh International Conference on Mobile Systems, Applications, and Services*, Kraków, Poland, June 22–25, 2009. New York, NY: ACM, pp. 179–192, 2009.

[13] P. Aghera, T. Rosing, D. Fang, and K. Patrick, "Poster abstract: Energy management in wireless healthcare systems," in *International Conference on Information Processing in Sensor Networks, 2009. IPSN 2009*, San Francisco, CA. Piscataway, NJ: IEEE, pp. 363–364, April 13–16, 2009.

[14] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making smartphones last longer with code offload," in *Proceedings of the Eighth International Conference on Mobile Systems, Applications, and Services, MobiSys'10*. New York, NY: ACM, pp. 49–62, 2010.

[15] S. Kang, J. Lee, H. Jang, H. Lee, Y. Lee, S. Park, T. Park, and J. Song, "SeeMon: Scalable and energy-efficient context monitoring framework for sensor-rich mobile environments," in *MobiSys'08: Proceeding of the Sixth International Conference on Mobile Systems, Applications, and Services,* Breckenridge, CO, June 17–20, 2008. New York, NY: ACM, pp. 267–280, 2008.

[16] A. Krause, M. Ihmig, E. Rankin, D. Leong, S. Gupta, D. Siewiorek, A. Smailagic, M. Deisher, and U. Sengupta, "Trading off prediction accuracy and power consumption for context-aware wearable computing," in *ISWC'05: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*. Washington, DC: IEEE Computer Society, Business Innovation Center Osaka, Osaka, Japan, pp. 20–26, October 18–21, 2005.

[17] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram, "Markov-optimal sensing policy for user state estimation in mobile devices," in *Proceedings of the First Annual International Conference on Mobile Computing, Applications, and Services—MobiCASE 2009,* San Diego, CA, October 26–29, 2009.

[18] M. Li, V. Rozgić, G. Thatte, L. Sangwon, B. A. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan, "Multimodal physical activity recognition by fusing temporal and cepstral information," IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18, 369–380, 2010.

[19] J. C. Sriram, M. Shin, T. Choudhury, and D. Kotz, "Activity-aware ECG-based patient authentication for remote health monitoring," in *ICMI-MLMI'09: Proceedings of the 2009 International Conference on Multimodal Interfaces*, Cambridge, MA. New York, NY: ACM, pp. 297–304, November 2–4, 2009.

[20] K. Patrick, W. G. Griswold, F. Raab, and S. S. Intille, "Health and the mobile phone," American Journal of Preventive Medicine, 35(2), 177–181, 2008.

[21] J. Ryder, B. Longstaff, S. Reddy, and D. Estrin, "Ambulation: A tool for monitoring mobility patterns over time using mobile phones," in *International Conference on Computational Science and Engineering, 2009. CSE'09*, vol. 4, Vancouver, BC. Piscataway, NJ: IEEE, pp. 927–931, August 29–31, 2009.

[22] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, BME-32(3), 230–236, 1985.

[23] The open source ECG toolbox (OSET). Available at: http://ee.sharif.edu/~ecg (accessed on January 17, 2016).

[24] Nokia energy profiler v1.2. Available at: http://m.en.softonic.com/app/nokia-energy-profiler/symbian (accessed on February 22, 2016).

[25] N. Agrawal, V. Prabhakaran, T. Wobber, J. D. Davis, M. Manasse, and R. Panigrahy, "Design tradeoffs for SSD performance," in *ATC'08: USENIX 2008 Annual Technical Conference*, Boston. Berkeley, CA: USENIX Association, pp. 57–70, June 22–27, 2008.

[26] G. Thatte, M. Li, S. Lee, B. A. Emken, M. Annavaram, S. Narayanan, D. Spruijt-Metz, and U. Mitra, "Optimal time-resource allocation for energy-efficient physical activity detection," IEEE Transactions on Signal Processing, 59(4), 1843–1857, 2011.

[27] D. P. Bertsekas, Problems with imperfect state information. Dynamic Programming and Optimal Control, 3rd ed. Belmont, MA: Athena Scientific, vol. 1, 2005, pp. 251–270.

[28] D. Lainiotis, "A class of upper bounds on probability of error for multihypotheses pattern recognition," IEEE Transactions on Information Theory, 15(6), 730–731, 1969.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, Bayesian decision theory. Pattern Classification, 2nd ed. New York, NY: Wiley-Interscience, 2001, pp. 20–82.

[30] R. T. Rockafellar and R. J. B. Wets, Horizon functions. Variational Analysis, 1st ed. New York, NY: Springer-Verlag, 1997, p. 95.

[31] R. E. Kalman, "A new approach to linear filtering and prediction problems," Transactions of the ASME—Journal of Basic Engineering, 82(Series D), 35–45, 1960.

[32] D.-S. Zois, M. Levorato, and U. Mitra, "Active classification for POMDPs: A Kalman-like state estimator," IEEE Transactions on Signal Processing, 62(23), 6209–6224, 2014.

[33] D.-S. Zois, "Active state tracking in heterogeneous sensor networks," Ph.D. dissertation, University of Southern California, Los Angeles, CA, August 2014.

[34] H. L. Van Trees and K. L. Bell, Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking. Piscataway, NJ: IEEE Press/Wiley-Interscience, 2007.

[35] G. Anastasi, M. Conti, M. D. Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," Ad Hoc Networks, 7(3), 537–568, 2009.

[36] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, "Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection, in *Proceedings of the Fifth European Conference on Wireless Sensor Networks (EWSN)*, Bologna, Italy. Berlin/Heidelberg: Springer-Verlag, pp. 17–33, February 2008.

[37] S. Balasubramanian, I. Elangovan, S. K. Jayaweera, and K. R. Namuduri, "Distributed and collaborative tracking for energy-constrained ad-hoc wireless sensor networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, Atlanta, GE. Piscataway, NJ: IEEE, pp. 1732–1737, March 21, 2004.

[38] P. Dutta, M. Grimmer, A. Arora, S. Bibyk, and D. Culler, "Design of a wireless sensor network platform for detecting rare, random, and ephemeral events," in *Proceedings of*

the Fourth ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN), Los Angeles, CA. Piscataway, NJ: IEEE, pp. 497–502, April 2005.

[39] Q. Cao, T. Abdelzaher, T. He, and J. Stankovic, "Towards optimal sleep scheduling in sensor networks for rare-event detection,"in *Proceedings of the Fourth ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN)*, Los Angeles, CA. Piscataway, NJ: IEEE, pp. 20–27, April 2005.

[40] H. Wang, H.-S. Choi, N. Agoulmine, M. J. Deen, and J. W. K. Hong, "Information-based energy efficient sensor selection in wireless body area networks," in *Proceedings of the IEEE International Conference on Communications (ICC)*, Kyoto, Japan. Piscataway, NJ: IEEE, pp. 1–6, June 2011.

[41] L. Stabellini and J. Zander, "Energy efficient detection of intermittent interference in wireless sensor networks," International Journal of Sensor Networks (IJSNet), 8(1), 27, 2010.

[42] M. Shin, P. Tsang, D. Kotz, and C. Cornelius, "DEAMON: Energy-efficient sensor monitoring, in *SECON'09: Proceedings of the Sixth Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*. New York, NY: ACM, pp. 565–573, 2009.

[43] A. Silberstein, R. Braynard, and J. Yang, "Constraint chaining: On energy-efficient continuous monitoring in sensor networks," in *SIGMOD'06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. New York, NY: ACM, pp. 157–168, 2006.

[44] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin, "Transmission scheduling for optimizing sensor network lifetime: A stochastic shortest path approach," IEEE Transactions on Signal Processing, 55(5), 2294–2309, 2007.

[45] J. L. Williams, J. W. Fisher, and A. S. Willsky, "Approximate dynamic programming for communication-constrained sensor network management," IEEE Transactions on Signal Processing, 55(8), 4300–4311, 2007.

[46] V. Krishnamurthy and D. Djonin, "Structured threshold policies for dynamic sensor scheduling—a partially observed Markov decision process approach," IEEE Transactions on Signal Processing, 55(10), 4938–4957, 2007.

[47] A. Seyedi and B. Sikdar, "Energy efficient transmission strategies for body sensor networks with energy harvesting," IEEE Transactions on Communications, 58(7), 2116–2126, 2010.

[48] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram, "Markov-optimal sensing policy for user state estimation in mobile devices," in *Proceedings of the Ninth ACM/ IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Stockholm, Sweden. pp. 268–278, April 12–16, 2010.

[49] Y. Wang, B. Krishnamachari, and M. Annavaram, "Semi-Markov state estimation and policy optimization for energy efficient mobile sensing," in *Proceedings of the Ninth Annual IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Seoul, Korea. pp. 533–541, June 2012.

[50] J. A. Fuemmeler, G. K. Atia, and V. V. Veeravalli, "Sleep control for tracking in sensor networks," IEEE Transactions on Signal Processing, 59(9), 4354–4366, 2011.

[51] G. K. Atia, V. V. Veeravalli, and J. A. Fuemmeler, "Sensor scheduling for energy-efficient target tracking in sensor networks," IEEE Transactions on Signal Processing, 59(10), 4923–4937, 2011.

[52] L. K. Au, A. A. T. Bui, M. A. Batalin, and W. J. Kaiser, "Energy-efficient context classification with dynamic sensor control," IEEE Transactions on Biomedical Circuits and Systems, 6(2), 167–178, 2012.

[53] M. A. Viredaz, L. S. Brakmo, and W. R. Hamburgen, "Energy management on handheld devices," Queue, 1(7), 44–52, 2003.

[54] E. Shih, P. Bahl, and M. J. Sinclair, "Wake on wireless: An event driven energy saving strategy for battery operated devices," in *MobiCom'02: Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking*. New York, NY: ACM, pp. 160–171, 2002.

[55] J. Sorber, N. Banerjee, M. D. Corner, and S. Rollins, "Turducken: Hierarchical power management for mobile devices," in *MobiSys'05: Proceedings of the Third International Conference on Mobile Systems, Applications, and Services*. New York, NY: ACM, pp. 261–274, 2005.

[56] L. Yan, L. Zhong, and N. K. Jha, "Energy comparison and optimization of wireless body-area network technologies," in *Proceedings of the ICST Second International Conference on Body Area Networks*. Brussels, Belgium: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, pp. 1–8, 2007.

[57] M. Shin, P. Tsang, D. Kotz, and C. Cornelius, "DEAMON: Energy-efficient sensor monitoring," in *IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, Rome, June 22–26, 2009.

[58] Y. Liu, B. Veeravalli, and S. Viswanathan, "Critical-path based low-energy scheduling algorithms for body area network systems," in *IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, Daegu, August 21–24, 2007.

[59] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *IMC'09: Proceedings of the Ninth ACM SIGCOMM Conference on Internet Measurement Conference*, Chicago, IL, November 4–6, 2009. New York, NY: ACM, pp. 280–293, 2009.

[60] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proceedings of the Eighth International Conference on Mobile Systems, Applications, and Services, MobiSys'10*, San Francisco, CA, June 15–18, 2010. New York, NY: ACM, pp. 209–222, 2010.

[61] A. J. Nicholson and B. D. Noble, "BreadCrumbs: Forecasting mobile connectivity," in *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, MobiCom'08*, San Francisco, CA, September 14–19, 2008. New York, NY: ACM, pp. 46–57, 2008.

[62] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely, "Energy-delay tradeoffs in smartphone applications," in *Proceedings of the Eighth International Conference on Mobile Systems, Applications, and Services*, San Francisco, CA, June 15–18, 2010. New York, NY: ACM, pp. 255–270, 2010.

9

MARKOV DECISION PROCESS FOR ADAPTIVE CONTROL OF DISTRIBUTED BODY SENSOR NETWORKS

Shuping Liu[1], Anand Panangadan[2], Ashit Talukder[3], and Cauligi S. Raghavendra[2]

[1] *eBay, Inc., San Jose, CA, USA*
[2] *Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA*
[3] *Information Access Division, National Institute of Standards and Technology, Gaithersburg, MD, USA*

9.1 INTRODUCTION

Wireless sensor networks are being increasingly used in scientific and defense applications. In these applications, multiple homogeneous or heterogeneous sensors make simultaneous measurements of their environment to obtain better estimates than is possible from one sensor alone. However, energy and computation capacity still remain the most significant bottlenecks in practical deployability of sensor networks. Limitations of battery capacity and the high energy cost of radio transmission necessitate energy-efficient communication and onboard processing to extend system lifetime in practically useable scenarios [1].

We consider a mobile, continuous health monitoring body sensor network where the size and ergonomic constraints severely limit the availability of power to battery sources that can be recharged only every few days or weeks. In this application,

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

a patient's vital signs are continuously measured by using multiple physiological and metabolic sensors attached to the patient's body (also called a "body sensor network"). Multiple sensors enable the subject's health status to be determined with higher accuracy (as in the case of using multiple accelerometers on different parts of the body to deduce the physical activity type [2]) or to reduce the effect of sensor variability through averaging; novel minimally invasive metabolite sensors, such as the interstitial fluid (ISF) alcohol sensor that we have studied, have significantly higher error rates than the corresponding invasive blood sampling-based sensors [3]. Unlike other sensor network applications, only a relatively few sensors (<5) are attached to a person. These sensors typically take measurements at a relatively low rate when the sensor readings are within normal limits, with higher sampling rates during critical periods. An important performance metric in this application is the lifetime of the body sensor network. The energy reserves within a sensor node get depleted with increasing amounts of sensing, communication, and computation. For instance, some physiological sensors have electromechanical components that make sensing an energy-intensive process [3]. Ideally, the system of sensors should operate for an extended period while maintaining an acceptable level of sensor performance.

Our body sensor network hardware system is described in Refs. [3, 4] and is based on the work performed at the Children's Hospital Los Angeles (CHLA) and the University of Southern California [3, 5]. In this system, multiple biomedical sensors can be attached to a human subject. The sensors include pulse oximeters to measure physiological parameters such as heart rate and blood oxygenation (noninvasive) and a novel sensor that estimates concentrations of metabolites of relatively small molecules (glucose, ethanol) from ISF samples collected through micropores created on the skin (a minimally invasive procedure). As the volume of ISF collected is very small (10 μl), the accuracy of metabolite measurements from one sensor is generally lower than measurements made from blood (the gold standard). In addition, due to issues with variability in the manufacturing process, the sensors have different noise levels. Therefore, multiple ISF sensors are used simultaneously to compensate for the inherent variation in measurement sensitivity between sensors. Each of these sensors is interfaced with a wireless node that provides its own energy source and the ability to communicate with the other sensors and to transmit sensor data. The sensor nodes can be independently controlled, and it is this control policy to operate the functioning of all sensor nodes in the body area network that is the focus of this work.

We describe in detail how to frame the coordinated sensing problem using the formalism of Markov decision processes (MDPs). The goal is to determine a *coordinated* sampling policy for each sensing node in the sensor network such that the overall system lifetime is extended without compromising the detection and monitoring of critical and/or life-threatening events. Specifically, the system must operate for a minimum specified length of time without every sensor running out of power while still maximizing the sensing rate during critical times. However, the MDP solution requires that the state of the distributed sensors is known at every control step. In a wireless health monitoring system, this implies radio communication of local state information between sensor nodes. As wireless communication is an energy-intensive process, we present an approximation method for executing the

MDP solution that does not require communication of state information at every step. In our approach, each sensor maintains an estimate of other sensor's internal state (i.e., its energy reserves). This estimate is used by the sensor to choose its sampling rate from the pre-computed solution. As the error in the estimate increases over time, sensor nodes periodically have to exchange their true local state via communication in order to reset their estimates. We present a new *value of information* (the information entropy in the estimate)-based communication scheme to reduce communications between sensors and thereby conserve energy. Communication broadcast only occurs when the expected value of information resulting from the communication exceeds a given threshold.

This chapter is organized as follows. In Section 9.2, we introduce MDPs and the rationale for using this formalism. In Section 9.3, we describe related work in energy conservation in sensor networks and MDPs. In Section 9.4, we describe the mathematical models for representing the sensor network as an MDP and the associated model functions. In Section 9.5, we discuss our approach for communication between sensor nodes to execute the coordinated sampling policy. Section 9.6 describes the solution to the communication and coordination problem. In Section 9.7, we present the detailed simulation results for global policy computation and lifetime analysis of this sensor system for different operational parameters of the system. In Section 9.8, we give our concluding remarks.

9.2 RATIONALE FOR MDP FORMULATION

An MDP is a discrete state space representation of the system being modeled with transitions between system states being defined stochastically with the Markov assumption (i.e., the probability of transitioning from the current state to another is dependent only on the current state and not on any of the previous states). The state transitions are influenced by actions that can be performed in a state. A reward function ascribes a value to each state that represents the desirability of reaching that state. The decision process is then to determine the optimal sequence of actions that will maximize the reward collected after starting from the initial state.

An MDP is a 4-tuple (S, A, P, R) where

- S is a finite set of states, in one of which the world exists.
- A is a set of actions that may be executed at any state.
- P is a probability function that defines how the state changes when an action is executed: $P : S \times A \times S \rightarrow [0,1]$. The probability of moving from state s to state s' after executing action $a \in A$ is denoted $p(s, a, s')$. The probability of moving to a state is dependent only on the current state (the Markov property).
- R is the reward function: $R : S \times A \rightarrow \boldsymbol{R}$. $R(s, a)$ is the real-valued reward for performing action a when in state s.

A *policy* is defined as a function that determines an action for every state $s \in S$.

The quality of a policy is the expected sum of future rewards. Future rewards are discounted to ensure that the expected sum of rewards converges to a finite value, that is, a reward obtained t steps in the future is worth $\gamma^t, 0 < \gamma < 1$, compared to receiving it in the current state. γ is called the discount factor. The *value* of a state s under policy π, denoted by $V^{\Pi}(s)$, is the expected sum of rewards obtained by following the policy π from s. The value function determines the action to be executed in state s under π:

$$\underset{a \in A}{\operatorname{argmax}} \left(R(s,a) + \gamma \sum_{s' \in S} p(s,a,s') V^{\Pi}(s') \right) \tag{9.1}$$

A policy is optimal if the value of every state under that policy is maximal. If all the model parameters are known, the optimal policy can be computed by solving Bellman's equations:

$$V(s) = \max_{a \in A} \left(\sum_{s' \in S} p(s,a,s') \left[R(s,a) + \gamma V(s') \right] \right) \tag{9.2}$$

Bellman's equations are solved using the value iteration algorithm. In this algorithm, the value function is initialized to arbitrary values $V_0(s)$. At iteration $k > 0$, the value function is updated:

$$V_k(s) = \max_{a \in A} \left(\sum_{s' \in S} p(s,a,s') \left[R(s,a) + \gamma V_{k-1}(s') \right] \right) \tag{9.3}$$

As $k \to \infty$, V_k converges to the optimal policy values.

The main advantage of the MDP formulation for our purpose is that the policy can be computed offline. Only the policy table needs to be stored in the network node. In this way, though the computational cost of computing a sophisticated policy offline may be high, only a simple table lookup is sufficient to execute the policy when the system is operational. This makes it suitable for implementation using the low-power processors typically found in embedded sensor network nodes.

For our problem, the MDP formulation represents the relevant features of the sensing system (energy consumption rates, expected changes in event criticality) as part of the overall stochastic model. The energy reserves, control time, and event criticality are represented as discrete variables. These variables define the state of the system at any time. In addition, the utility of coordinated sampling and the penalty for running out of energy are quantified into a reward associated with each state. The objective is for the system to operate for a minimum specified length of time without the sensors running out of power while optimizing the reward function for health monitoring. The model is "solved" to determine a policy that specifies the sampling rate for each sensor for every possible state of the system (i.e., for different amounts of energy reserves at a sensor node). The policy obtained in this manner is optimal under the assumptions of the underlying model.

The general modeling approach is illustrated in Figure 9.1. The states of the system are shown distributed on a grid. States from left to right represent successive

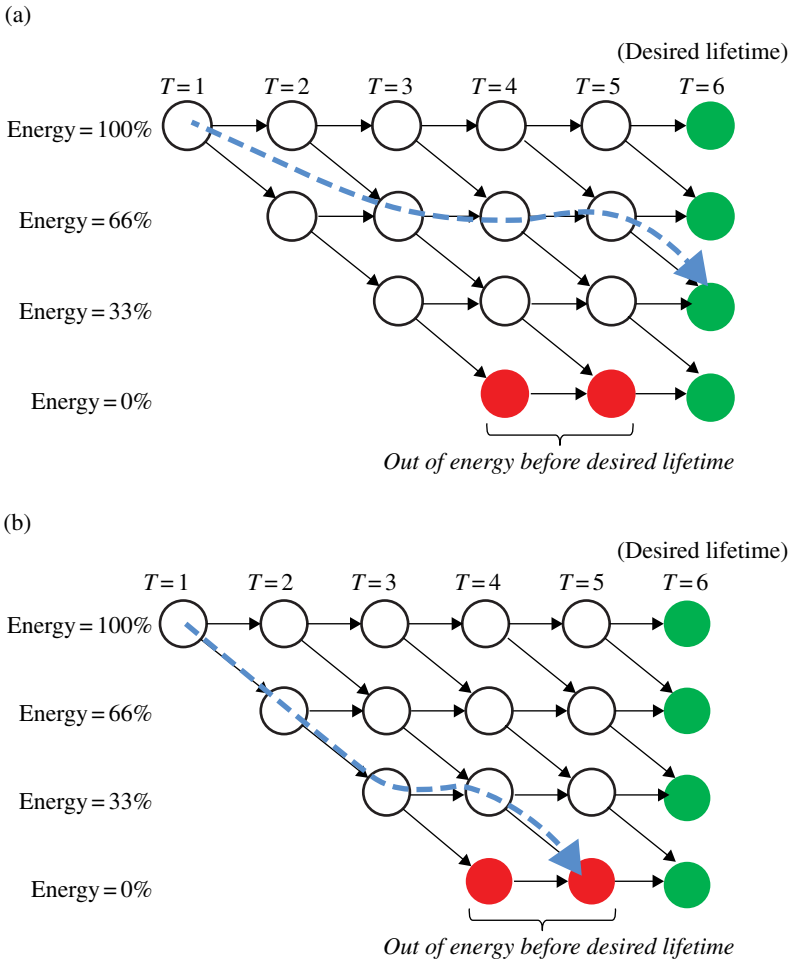


FIGURE 9.1 (a) An execution where the sequence of states (dashed line) ended in a desirable end state (system did not run out of energy at the end of the desired lifetime). (b) An execution that ended in an undesirable state—the system ran out of energy in less than the desired lifetime.

time-steps, while states from top to bottom represent energy being consumed. The arrows represent state transitions. Note that at every time-step, each state can transition to either a state with the same energy reserve (e.g., due to low sensing) or to a state with smaller energy reserve (e.g., due to sampling at a high rate). The goal is to come up with a set of actions that will lead to a desirable end state (as shown in Fig. 9.1a) where the system has not run out of power before the end of the desired lifetime. An undesirable execution is one where the system reaches a state with no energy before the end of the desired lifetime (as shown in Fig. 9.1b). The optimal execution is one where the system runs out of power exactly at the end of the desired

lifetime since this represents that all the energy was used for sensing (and leftover energy at the end represents unused sensing resources). The challenge is that the state sequence cannot be directly controlled—only the actions (sampling rate) can be controlled, which affects the state transition in a probabilistic manner.

9.3 RELATED WORK

There has been extensive research in the past decade to improve energy efficiency in embedded sensor networks at multiple levels, including at the hardware platform, operating system, network protocols, and application software [1]. Chandrakasan and Brodersen [6] developed a low-power complementary metal oxide semiconductor (CMOS) controller for each sensor. Raghunathan et al. [7] incorporates energy awareness into the task scheduling process at the operating system level. As the active state of the radio consumes a relatively large amount of energy, managing sleep cycles of a collection of nodes is a viable means of energy conservation. Ye et al. [8] employs a periodic cycle to coordinate the sleeping and waking processes. Each sensor node sleeps for a fixed duration and then wakes to listen for transmissions. Energy conservation through sleep cycles assumes that the environment is densely populated with sensors and needs to be sampled only by a subset of the sensors at a time. In our work, we consider the case where all sensors sample the environment continuously. Adaptive sampling has also been used as a means of energy conservation. Willett et al. [9] employ a node as a fusion center to collect an initial estimate of sensor measurements from a small set of nodes and then "backcasts" the estimate to selectively activate additional sensor nodes. However, there has been relatively little work on incorporating resource management into wireless health monitoring systems. Aghera et al. [10] consider energy conservation during task allotment in a wireless healthcare system. Krause at al. [11] explicitly trade off energy consumption for activity classification accuracy in a accelerometer-based sensor system.

MDPs have been used previously as a predictive model and for the coordination between agents. In our prior work, we have used MDPs to adapt the sampling rates of a single sensor to energy expenditure [5]. In the current work, we extend the approach to a system with multiple correlated sensors and introduce communication control in the same framework. Wang et al. use constrained MDPs to efficiently estimate user state in a wireless device [12]. Wang et al. also model the relationship between energy expenditure and state estimation accuracy as a Markov chain [13]. However, these formulations do not explicitly model system time. Zilberstein et al. [14] formulate the coordinated sensing problem for lunar rovers using transition-independent MDPs. As in our work, the ability to compute the entire policy offline and thus make only limited use of the computing power onboard the rover was listed as one of the advantages of this scheme. Becker et al. [15] analyze the complexity of this type of MDP. Considering that the sensors in our application are distributed without knowledge of each other's states, distributed partially observable MDPs (distributed POMDP) are ideally suited to plan such policies [16, 17]. However,

the problem of finding the optimal joint policy for general distributed POMDPs is NEXP-complete [18]. Therefore, we use completely observable MDPs during policy generation (assuming that all sensors can make perfect observations), but during policy execution under partial observability, the sensors communicate to exchange state information.

MDP-based frameworks have also been used to determine a communication protocol to coordinate the actions of distributed agents. Xuan et al. [19] consider two separate MDPs that use a common global utility function and model communication as an explicit action that incurs a cost. Their approach is applicable only if the transition models are independent. Communication using the framework of POMDP has also been studied [20, 21]. However, in distributed POMDPs, the central planner must reason explicitly about the possible observations of all agents when generating a policy for one agent [22]. This reasoning about observations makes distributed POMDPs much more complex than distributed MDPs. These approaches model communication as an explicit or implicit action during the construction of the global policy. This significantly increases the size of the action space. In our work, we do not consider communication during the policy building phase. Communication decisions are made only during policy execution.

Wireless sensor networks are being increasingly used for human health monitoring (often called "body sensor networks" [23]). CodeBlue is a wireless network designed for use in emergency medical care [24]. The authors have integrated physiological and metabolic sensors into a wireless network [25]. Joshi et al. suggested a mobile cardiac outpatient telemetry (MCOT) system [26], which does not require activation by the patient, and is mobile and flexible. MCOT system only monitors the cardiac symptom of patients and provides a service for communicating the information to doctors through cardionet. We are interested in monitoring patients with coordination among multiple sensors for supporting mobile patient monitoring.

9.4 PROBLEM STATEMENT, ASSUMPTIONS, AND APPROACH

9.4.1 Problem Statement

We consider a system where multiple sensors observe the same phenomenon and make measurements. Each sensor has its own dedicated energy source. Increasing the sampling rate increases the rate of energy consumption. (In a physical system, this could be because the sensor operation itself consumes significant amounts of power or that the larger amounts of data resulting from higher sampling rates have to be transmitted using wireless transmission.) The capacity of the energy source is limited, that is, sampling at the highest rate continuously will result in exhausting the energy source before the desired lifetime.

The sensors can communicate with each other but communication has an energy cost. For simplicity, in this work a communication step is defined to be exchange of all local information among all the sensors (fully connected network). The system should remain in operation for some predefined duration, that is, the sensor nodes

should not all run out of power before this time. The system can adapt the sensor sampling rates to changes in the criticality of the subject's heath condition. The definition of criticality will be application dependent. For instance, in a diabetic monitoring system where sensors estimate blood glucose concentration, the criticality may be defined to be proportional to the deviation of the concentration from normal levels. It is expected that more frequent measurements will be taken when the event criticality is high to reduce measurement uncertainty while less frequent measurements will be sufficient at other times.

The problem is to determine the sampling rates of the individual sensors such that the health condition is monitored with high accuracy, but the system remains in operation until the desired lifetime. If the sampling rates are too high, then the sensors may run out of power before the desired lifetime. On the other hand, if they are too low, then the phenomenon is sampled suboptimally.

9.4.2 Assumptions

We assume that the sensors make independent but correlated measurements with Gaussian random errors. Under this assumption, the error variance in the measurement estimate can be reduced by averaging the individual sensor readings. This corresponds to increasing the sampling rate of the sensors.

We assume that the expected change in criticality of the health status can be modeled as a Markov process. In our work, the Markov model is used as a simple model of the time-varying health status of the subject. We also assume that the criticality is provided to the system—either from an external source or computed from the sensor measurements.

We have used a discrete state space MDP for modeling the problem. Thus, it is necessary to discretize the continuous variables (i.e., energy reserve of the sensors). The size of the state space of the MDP increases with how finely the variables are resolved. The state space also increases exponentially with the number of sensors in the system. In body sensor networks, the number of biomedical sensors is relatively few, and hence the resulting MDP can be solved in a reasonable amount of time.

9.4.3 Approach

We model the problem described before as an MDP. In our MDP formulation, the state represents the energy reserves at every sensor at a given time and the criticality of the subject's heath condition. The state changes when there is a change in the energy reserve or when the data criticality changes. Therefore, the energy consumption rates and the expected changes in data criticality are modeled as stochastic transitions between system states. The actions correspond to sensor sampling rates. The reward in each state quantifies the utility of sampling at particular rates and the penalty of running out of power. The MDP formulation enables us to compute the optimal policy, which specifies the optimal sampling rate for every sensor at every possible state of the system. This optimal policy is then stored within each sensor node for execution after deployment.

This state space formulation assumes that the internal state (current energy reserve) of all the sensors is "completely observable" by a sensor at every control step, that is, every sensing node knows the energy reserve of every other node at any time. But during execution, the world is only "partially observable," that is, one sensor does not know the local status (consumed energy) of the other sensors. In a distributed network, such a *global* policy can be executed only if the nodes exchange state information before every control step. In a wireless network, such exchanges are energy intensive. We have formulated a method that is less communication intensive for executing the optimal global policy. In our method, every sensor node maintains an estimate of the other sensors' internal states, utilizing the same stochastic model used for policy creation. A sensor uses this estimated global state to choose the action from the MDP global policy table.

Over time, the accuracy of these local estimates will decrease. At every control step, a sensor node computes the information entropy in the local estimate. If this entropy exceeds a threshold, the sensor node triggers an exchange of the true state information with all other sensors. Thus, communication is initiated only if the expected *value of information* that is to be received is above a predefined limit. After a communication step, the sensor nodes learn the true global state and the entropy of their local estimates reduces to the minimum value. This process is illustrated in Figure 9.2.

Note that the entire MDP policy is computed offline before node deployment using the stochastic model. Only the resulting policy is stored in the memory of each node. At every control step, the sensor estimates the current state of the system and then

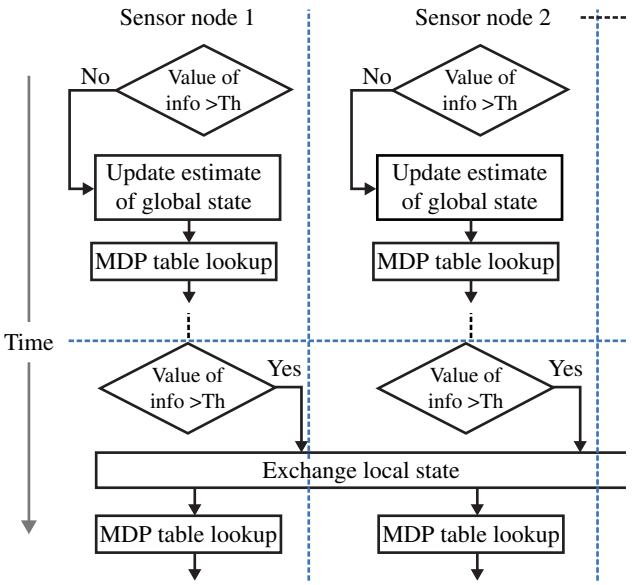


FIGURE 9.2 Interlacing of maintaining local state estimates and learning true global state via communication.

performs a table lookup of the pre-computed policy. Thus, the computational cost of executing the control policy is minimal. (The computational cost of calculating the optimal policy is substantial and increases exponentially with the number of sensors.)

9.5 MDP MODEL FOR MULTIPLE SENSOR NODES

We now describe how the problem of coordinated sensing in a resource-constrained sensor network can be formulated as an MDP. We discretize all real-world features since the MDP formulation we use requires a discrete state space. Let N denote the number of sensors that are to coordinate with each other. The *local state* of node N_i is represented by the state vector (t, h_i, e_i). $t \in \{1,2,\ldots,T\}$ indicates the number of control steps completed since the initial time. T corresponds to the guaranteed lifetime desired from the joint system. $h_i \in \{1,2,\ldots,H\}$ is a measure of the criticality of the sensor readings. This is application dependent. For instance, this could correspond to the health condition of a patient in a human health monitoring application. $e_i \in \{1,2,\ldots,E\}$ is the amount of energy consumed. We assume that the event criticality can be estimated at each sensor node based on its local sensor measurements. Therefore, it is available to each sensor node, that is, $h_i = h, i = 1,2,\ldots,N$. Representing the state using components each representing an independent entity is called a feature-space representation.

The *global state* is the joint local states of all the sensors, $(S_1, S_2, \ldots, S_N)$. Let S denote the finite set of all possible global states. The joint action space A is the action concurrently executed by all sensors, $A = A_1 \times A_2 \times \cdots \times A_N$ where A_i is the action space of sensor node N_i and A denotes the set of all possible sensor sampling rates.

P is the transition probability function defining how the global state changes when a joint action is executed, $P : S \times A \times S \to [0,1]$. The probability of moving from state s_i to state s_j after taking action a is denoted by $p(s_i, a, s_j) = p((t_i, h_i, \boldsymbol{e_i}), (a_1, a_2, \ldots, a_N), (t_j, h_j, \boldsymbol{e_j}))$ where $\boldsymbol{e_i} = (e_{1,i}, e_{2,i}, \ldots, e_{N,i})$ and $\boldsymbol{e_j} = (e_{1,j}, e_{2,j}, \ldots, e_{N,j})$. The increase in control step, change in event criticality, and fall in energy reserves are independent, and hence we can define

$$p\left(s_i, a, s_j\right) = p_T\left(t_i, t_j\right) p_H\left(h_i, h_j\right) \prod_{k=1}^{N} p_E\left(e_{k,i}, a_k, e_{k,j}\right) \tag{9.4}$$

We define the component transition functions as follows:

$$p_{T\left(t_i, t_j\right)} = \begin{cases} 1, & \text{if } t_i = t_j = T \\ 1, & \text{if } t_j = t_i + 1 \\ 0, & \text{otherwise} \end{cases} \tag{9.5}$$

$$p_E\left(e_i, a, e_j\right) = \begin{cases} 1, & \text{if } e_i = E \\ p_P\left(a\right), & \text{if } e_j = e_i \\ 1 - p_P\left(a\right), & \text{if } e_j = e_i + 1 \\ 0, & \text{otherwise} \end{cases} \tag{9.6}$$

The rate at which energy is consumed by a sensor is dependent on the sampling rate (action) and modeled with probability $p_E(a)$. The energy consumption rate increases with the sampling rate.

$$p_{H(h_i,h_j)} = \begin{cases} p_H, & \text{if } i = j \\ 2p_H^{\text{change}}, & \text{if } h_i = 1, h_j = 2 \text{ or } h_i = H, h_j = H-1 \\ p_H^{\text{change}}, & \text{if } |h_i - h_j| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9.7}$$

p_H and p_H^{change} are probabilities that model the change in event criticality (the two values are dependent since the sum of all transition probabilities out of a state must sum to 1). These component transitions are illustrated in Figure 9.3 and the evolution of the full local state is shown in Figure 9.4.

$R = R(s,a) = R((t,h,e_1,e_2,\ldots,e_N),(a_1,a_2,\ldots,a_N))$ is the reward function, and it depends only on the sensor sampling rates and the system state (including event criticality). Intuitively, the sampling rate should be higher during critical events. There

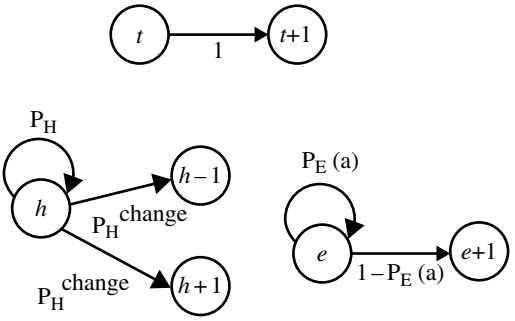


FIGURE 9.3 Component transitions of the MDP model; "t" represents the evolution of time, "h" the evolution of event criticality, and "e" the local (energy) resources at a node.

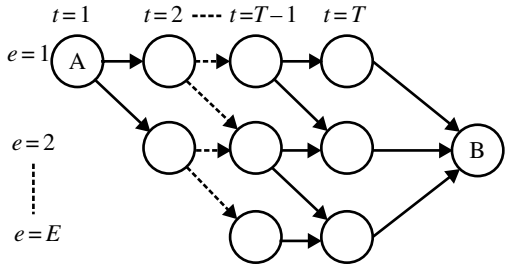


FIGURE 9.4 Evolution of the local state at a sensor node. State labeled "A" is the initial state ($t = 1$ and full energy reserve), and state labeled "B" is the terminal state ($t = T$). At every control step, the time element increases by 1, and the energy element increases stochastically. The event criticality element is not shown for clarity.

is a penalty, R_{powerout}, if the system (i.e., *all* sensor nodes) runs out of power before the desired lifetime. The reward function is defined as follows:

$$R\left(\left(t,h,e_1,e_2,\ldots,e_N\right),\left(a_1,a_2,\ldots,a_N\right)\right)=\begin{cases}R_{\text{powerout}}, & \text{if } t<T, e_i=E, i=1,2,\ldots,N\\ k\times\sum_{i=1,e_i<E}^{N} a_i\times h, & \text{otherwise}\end{cases} \tag{9.8}$$

If multiple sensors are able to sense simultaneously, then the reward is proportional to the sum of the sampling rates, that is, the reward is inversely proportional to the expected variance in the fused estimate. This is obtained from the assumption that successive sensor measurements are independent and that the true measurement is corrupted by zero-mean Gaussian noise. The formulation given in text holds true when the sensors have the same error variance and zero covariance. This term can be modified to reflect unequal error variances and covariances (e.g., using a Kalman filter formulation).

The global policy is obtained by solving this MDP through the value iteration or policy iteration algorithms.

Note that the entire MDP policy is computed offline before node deployment using the stochastic model. Only the resulting policy is stored in the memory of each node. At every control step, the sensor estimates the current state of the system and then performs a table lookup of the pre-computed policy. Thus, the computational cost of executing the control policy is minimal. (The computational cost of calculating the optimal policy is substantial and increases exponentially with the size of the state space.)

The standard formulation of the algorithms used to solve an MDP involves an explicit enumeration of the complete state space and transition table. The size of the state space is proportional to how finely the real-world features (time, energy reserves, event criticality, sampling rates) are discretized. In a multiple node setting, the size of the state space also increases exponentially with the number of sensor nodes. The large state space increases both the time required to compute the optimal policy (this is done offline) and the space required to store the resulting optimal policy within the embedded sensor nodes. Efficient representation of MDPs and computation of approximate policies is an active area of research [27, 28]. For instance, the computation efficiency can be increased when there are conditional independencies in the model [28]. The policy table may also be represented more efficiently than explicitly enumerating every state [29]. However in our current work, we have only implemented the standard formulation of the value iteration algorithm that requires an explicit enumeration of the state space. This restricts the maximum number of sensor nodes that can be modeled (though this is sufficient to model all the different sensors in our health monitoring system).

9.6 COMMUNICATION

We have formulated a method to decide when to communicate that makes use of the inherent ambiguity in set of possible global states at a control step. If the level of ambiguity is over a predefined threshold, then the sensor nodes exchange information

to identify the true global state (no ambiguity). We use the information theoretic measure of Shannon entropy to quantify this ambiguity. The entropy is least when the exact global state is known and maximum when the global state may be any of the possible states with equal probability. We now express these concepts in terms of the MDP.

Let t denote the control step when the decision to communicate has to be made by node N_i. Denote by subscript $-i$ the set of parameters of all nodes other than N_i (e.g., $e_{t,-i}$ denotes the set of energy reserves of all nodes other than N_i at time t). Let the global state at $t-1$ be precisely known to be $s_{t-1} = (t, h, e_{t-1,i}, e_{t-1,-i})$ and the action executed at $t-1$ be $a_{t-1} = (a_{t-1,i}, a_{t-1,-i})$. At control step t, N_i only knows its local energy reserve, $e_{t,i}$ (local state) but not that of the other nodes. (For simplicity of notation, we assume that the event criticality, h, does not change.) Let e_{-i} denote an arbitrary vector with components $(e_1, e_2, \ldots, e_{i-1}, e_{i+1}, \ldots, e_N)$. The set of possible states at t contains those states that have a nonzero transition from the state at $t-1$:

$$S_t = \left\{ \left(t, h, e_{t,i}, e_{-i}\right) \middle| p_E\left(e_{t-1,j}, a_{t-1,j}, e_j\right) > 0, \quad \forall j \neq i \right\} \tag{9.9}$$

Here, $p_E(a)$ is the transition probability function. The probability of the global state being a particular state $s \in S_t$, with components $s = (t, h, e_{t,i}, e_{-i})$ is given by

$$Pr_t(s) = \prod_{j=1, j\neq i}^{N} p_E\left(e_{t-1,j}, a_{t-1,j}, e_j\right) \tag{9.10}$$

Note that $Pr_t(s)$ is dependent only on the transition probability and the action taken in the previous step. We desire to calculate the ambiguity in the set of possible states after the current control-step, that is, at $t+1$. Let S_{t+1} denote this set. S_{t+1} contains all the states that can be reached from a state in S_t by taking the action specified by the policy. Let $\pi(s)$ denote the action (sampling rates) specified by the policy at a state s. Then, S_{t+1} is given by

$$S_{t+1} = \left\{ s \middle| \exists s' \in S_t, p\left(s', \pi\left(s'\right), s\right) > 0 \right\} \tag{9.11}$$

The probability of the global state at $t+1$ being a particular state $s \in S_{t+1}$ is given by

$$Pr_{t+1}(s) = \sum_{s' \in S_t} Pr_t\left(s'\right) p\left(s', \pi\left(s'\right), s\right) \tag{9.12}$$

Note that the distribution of states at $t+1$ is dependent on the policy, π. We use the information entropy of this distribution as the value of communication, V_t, and is given by

$$V_t = -\sum_{s \in S_{t+1}} Pr_{t+1}(s) \log\left(Pr_{t+1}(s)\right) \tag{9.13}$$

If V_t exceeds a predefined threshold (determined empirically in our current work), the sensor node triggers a communication step, which leads to exchange of local state information and exact knowledge of the global state at control step t.

Each communication action incurs a cost, which is also modeled stochastically with a probability parameter p_c. For each communication action, the energy reserve of the sensor node decreases by 1 with probability p_c and remains the same with probability of $1 - p_c$. The average energy consumed at each communication step is p_c.

9.7 SIMULATION RESULTS

We first describe an existing body sensor network from which we derive the simulation parameters such as the various transition probabilities (energy consumption rate), relative amounts of rewards/penalties, and communication cost. We then show simulations that demonstrate how the coordinated sensing policy reacts to both decreasing energy reserves and event criticality. We next report the results of the simulation experiments that demonstrate the effect of communication between sensor nodes on policy execution. After that, we evaluate the effect of changing an MDP parameter on the computed policy. We then evaluate the sensitivity of the policy to differences in the stochastic model parameters used during policy computation and policy execution. We finally compare our MDP-based coordinated sensing method with other approaches.

We use two metrics of system performance. Note that as the policy is stochastic, the metrics are expected values (obtained in simulation experiments by averaging the results from several policy executions). The first is the *system energy outage* percentage, which is defined as the proportion of policy executions that ended with *all* sensors running out of power before the desired lifetime (T in the MDP model). The second metric is the *system lifetime,* which is defined as the expected number of control steps that the system is in operation (at least one of the sensors has power).

9.7.1 Simulation Parameters

The simulation is based on a body sensor network developed at the CHLA. The system is designed to collect data from multiple sensors attached to the subject's body (noninvasive pulse oximeters and minimally invasive metabolite sensors) via wireless links. For purposes of simulating the MDP-based policy, we will consider the energy requirements of the metabolite sensor. This sensor collects ISF collected from the epidermis of the skin of the forearm [3]. The fluid is extracted using a miniature electromechanical vacuum pump (CTS Diaphragm Pump, E107-12-090). Multiple ISF sensors are used simultaneously to compensate for the inherent variation in measurement sensitivity between sensors.

Each sensor is interfaced with a low-power microcontroller (MSP430, Texas Instruments, Dallas, equipped with a CPU at 4–8 MHz) and a custom frequency agile and range agile radio circuit design capable of operating at the 900 MHz and 2.4 GHz radio frequency (RF) bands. The MDP-based coordinated sensing algorithms are to be implemented on the microcontroller. The sensor, transceiver system, and interface boards are packaged into a unit that can be attached to the subject with minimal

(a) (b)

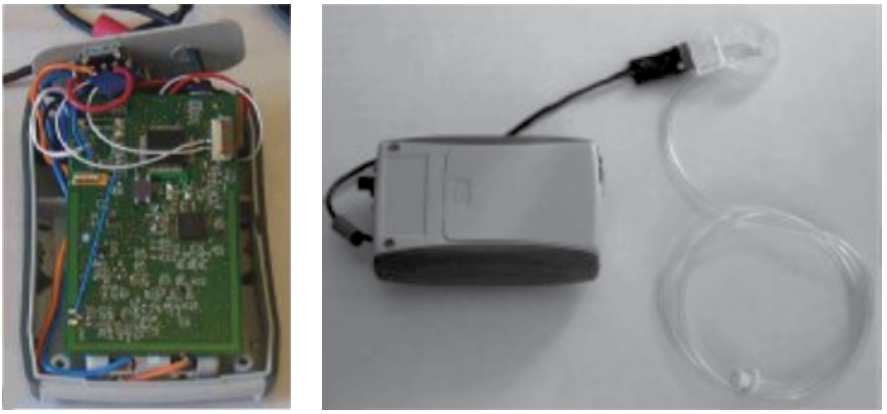


FIGURE 9.5 (a) Internal circuit of interstitial fluid (ISF) sensor containing microcontroller and radio transceivers and (b) ISF alcohol sensor (the tube to the vacuum pump is not connected).

TABLE 9.1 Battery Capacity and Energy Consumption Rates of the Sensor Components

Component	Energy Capacity or Consumption Rate
Battery capacity (CR2354)	560 mAh @ 3 V = 6048 J
Sleep mode of sensor node	10 mA @ 3 V = 0.03 W
Short-range radio	35 mA @ 3 V = 0.105 W
ISF pump	230 mA @ 9 V × 5 s = 10.35 J

hindrance to daily activities. Figure 9.5 shows photographs of the ISF alcohol sensor and the internal circuitry of the microcontroller interface.

The energy capacity of the battery used for the sensors and the energy consumption rate of the system components are shown in Table 9.1. The vacuum pump operates independently of the sensor (for periods of 5 s when the pressure falls below a predefined threshold) but shares the battery with the sensor.

In this application, the event criticality is proportional to the estimated metabolite concentration, that is, it is important to estimate the concentration with higher accuracy as the concentration rises above the nonzero level. However, in our simulation experiments, we predefine the criticality level over time as the data from the sensors is not simulated.

Each action level is defined by a period of time of operating the sensor followed by the sensor entering a sleep mode, that is, the duty cycle of the sensor varies for different action levels. The expected energy consumption rate can then be computed for each action level from the values in Table 9.1.

The expected energy consumption rates are then modeled as probabilities ($p_E(a)$) for use in the MDP model. Let $e'(a)$ denote the rate of energy expenditure per unit

time when performing action a, and let ΔE denote the unit of energy used for discretization. We model $p_E(a)$ (probability that the energy level does not change on performing action a) as percentage of the number of control steps where the discrete energy level did not change while taking constant action a to the total number of control steps, T:

$$p_E(a) \cong \frac{T - \left(e'(a)T/\Delta E\right) - 1}{T} = 1 - \frac{\left(e'(a)T/\Delta E\right) - 1}{T}$$

Taking action $a = 1$ as example, where the energy consumed in each time step for action 1 is

$$E(a = 1) = 67.725$$

Figure 9.6 shows the energy evolving process for $a = 1$ in both continuous energy space and discretized energy space. Each dot (solid or outlined) represents the time when action 1 is taking. Each stairstep is one discretized energy level. When taking action 1 beginning with full-charged battery, the total time steps with both same energy level and the changed energy level are 89, and the total time steps with energy level changing is 19 (Fig. 9.6). So

$$p_E(1) = \frac{89 - 19}{89} = 0.7865$$

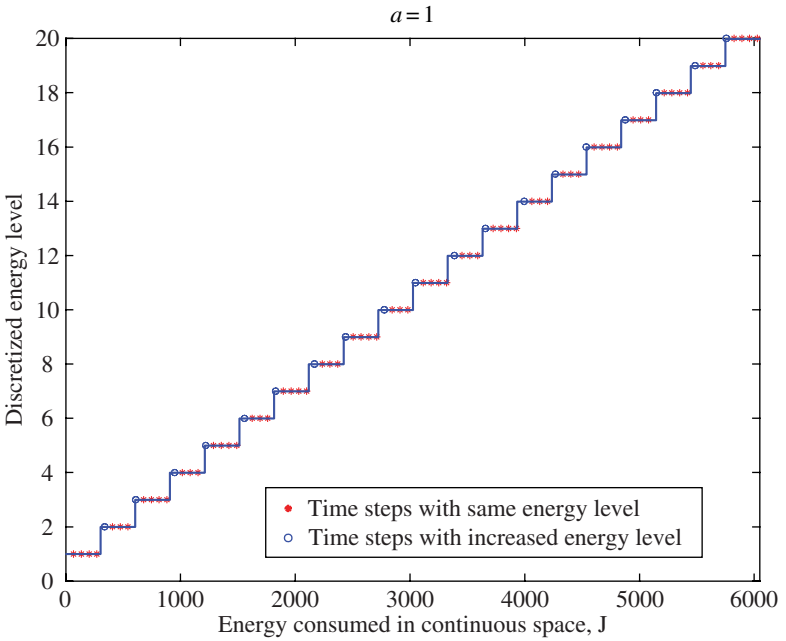


FIGURE 9.6 The energy evolving process in continuous energy space and discretized energy space.

TABLE 9.2 Model Parameters Used in Simulations

Number of Sensors	T	H	E	A	Number of States
2	48	10	20	4	192,000
3	40	10	10	2	400,000
5	14	4	4	2	57,344

$p_E(a)$ is a function of action a and decreases with the sampling rate (action), which means higher sampling rates will increase the chance to increase the energy level for each time step. The reason is that when action increases, the interval between two consecutive time steps increases.

We also set the probability of the data criticality level remaining constant to $p_H = 0.75$.

The number of states in the discrete MDP model is proportional to how finely time, criticality, and energy reserves are discretized and the state space exponentially increases with the number of sensors: $|S| = T \times H \times E^N$ where T is the number of control steps, H is the number of criticality levels, E is the number of energy levels, and N is the number of sensors. The size of the state space limits the size of the problems that can be solved using the discrete MDP approach. For instance, in the two-sensor case, we model $T = 48$ control steps, which in our alcohol sensing application corresponds to controlling the sensors every 30 min for a system lifetime of 24 h. Table 9.2 lists the parameters that are used in the simulation experiments.

9.7.2 Adapting to Energy Reserves and Data Criticality

We evaluated the policy over 10,000 executions using a constant event criticality profile to study how the sampling rates (actions of the MDP policy) vary over the lifetime of the system. The sampling rates are shown in Figure 9.7. At the early stages of the execution, the sampling rates are low indicating that the policy is conserving energy for an expected increase in criticality. However, toward the end of the desired lifetime, the sampling rate increases in order to measure the event with higher fidelity as the risk of running out of power decreases. This illustrates that the optimal policy varies from being conservative to more aggressive depending on the age of the system relative to the desired lifetime.

We next evaluated the policy using a time-varying event criticality profile to study how the sampling rates (actions of the MDP policy) vary with event criticality. Figure 9.8 shows the change in sampling rate over time with changes in event criticality for 2-sensor BASN systems. The sampling rate increases during periods of high criticality while the policy still ensures that the system does not run out of power.

9.7.3 Communication

In order to demonstrate the impact of communication on policy execution, we executed the policy under two extremes of communication: full or no exchange of state information. In the full-exchange case, the sensors communicate without incurring

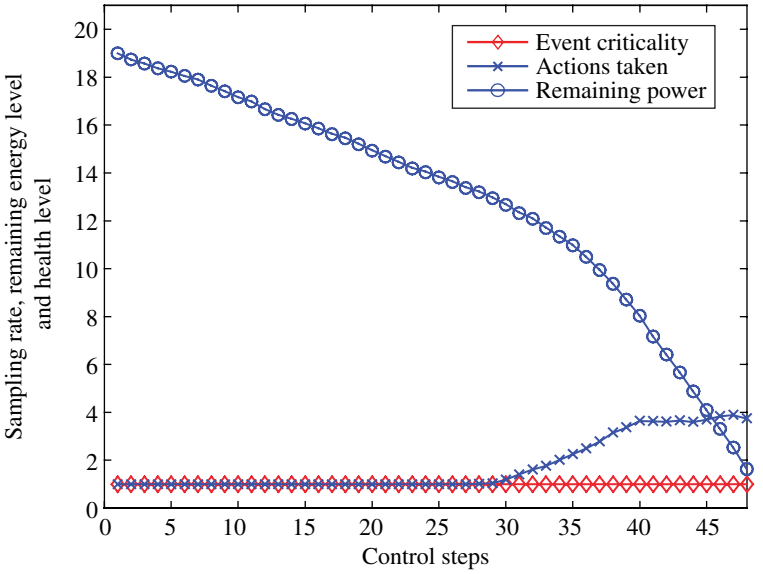


FIGURE 9.7 Sampling rate over time with constant event criticality.

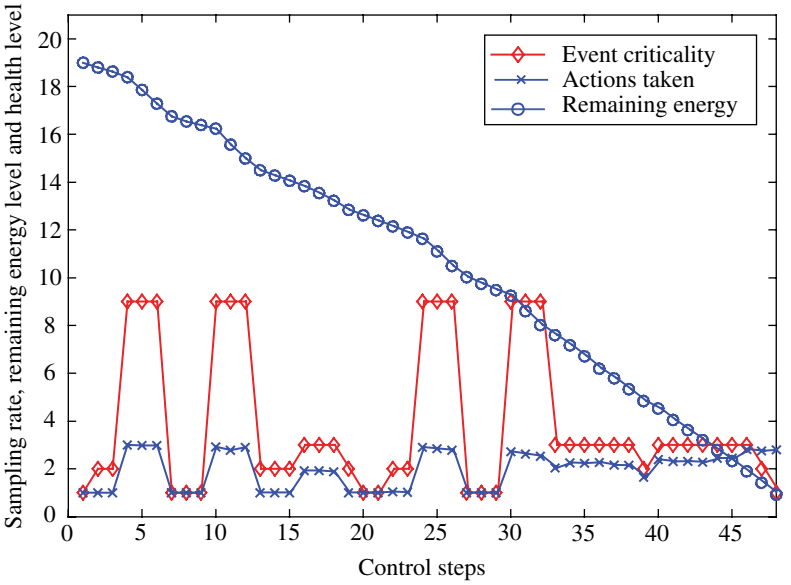


FIGURE 9.8 Change in sampling rate over time with changes in event criticality.

an energy cost before every control step to learn the full state. In the no-communication case, each sensor relies on a stochastic model of the other sensor's performance in lieu of the true state. Figure 9.9 shows the probability of the system running out of power under these two cases. The number of sensor nodes in the system varies from

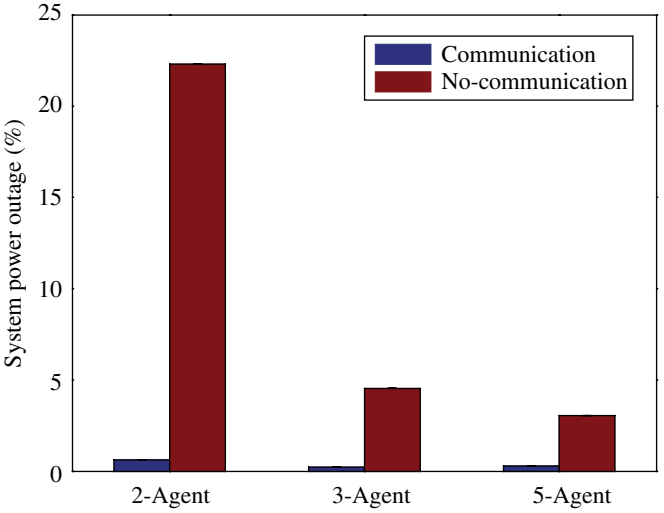


FIGURE 9.9 Probability of system running out of power under the full-communication and no-communication case.

2 to 5, with all the sensors having the same error covariance. We first quantify the effect of the full- and no-communication schemes on the execution of the MDP policy. The results are averaged from 10,000 simulation runs. As expected, full exchange of state information enables the execution of the optimal policy, and hence the system has a longer lifetime.

We next demonstrate how the threshold of communication affects the system's lifetime and utility of information extracted from the sensor system. The communication threshold represents the minimum uncertainty in the global state estimate required to trigger an exchange of local state information between nodes via communication. p_c represents the cost (probability that energy reserves keeps the same) of a communication step. Higher values of p_c corresponds to lower communication cost. Figure 9.10 shows the change in the probability of system power outage with variation of the communication threshold for $N = 2$ cases. Figures 9.11 and 9.12 show the corresponding average utilities and system lifetime. These figures show the trade-off between the energy cost of communication and the amount of information available from communication. Increased communication leads to better execution of the optimal policy, but it also incurs a higher energy cost from communication. On the one hand, when the threshold is low, the number of communication steps increases and this leads to a more faithful execution of the optimal policy, but the high energy cost of communication offsets the benefit of executing the optimal policy. Thus the system power outage increases. On the other hand, when the communication threshold is high, communication reduces and system performance begins to reflect the no-communication case. When the communication threshold is higher than some value (~2 in the two sensor experiments shown in Fig. 9.10), none of the sensors communicate and execute the global policy only using the local estimates of the

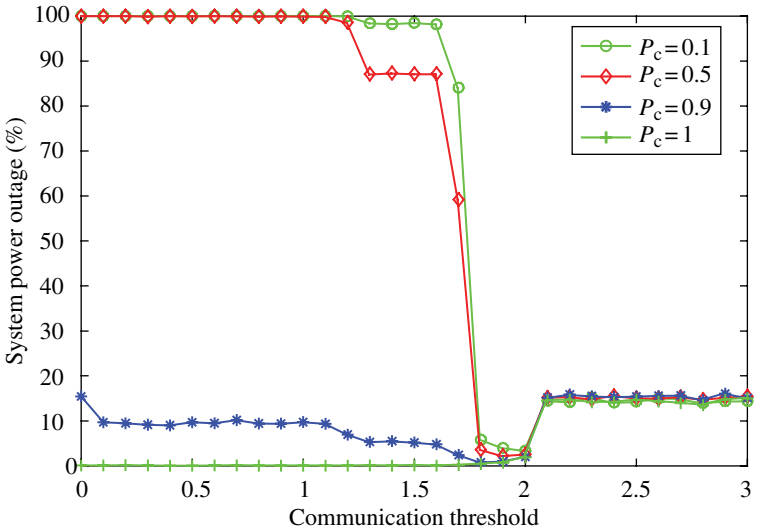


FIGURE 9.10 Effect of varying the threshold of communication on the system power outage ($N=2$).

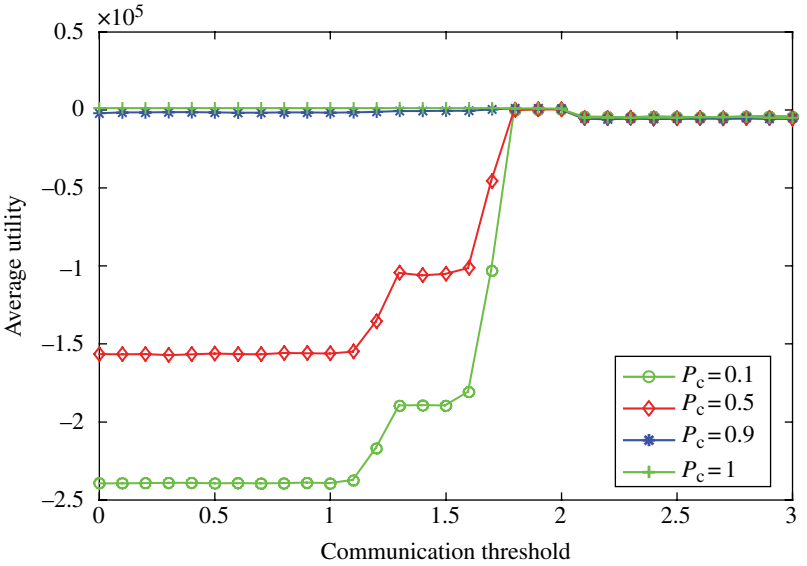


FIGURE 9.11 Effect of varying the threshold of communication on the average utility of executing the MDP policy ($N=2$).

global states. Beyond this communication threshold point, the performance of the system is the same for different values of p_c. Thus the communication threshold of approximately 2 represents the point where the system exhibits the smallest probability of running out of power before the desired lifetime, that is, the best

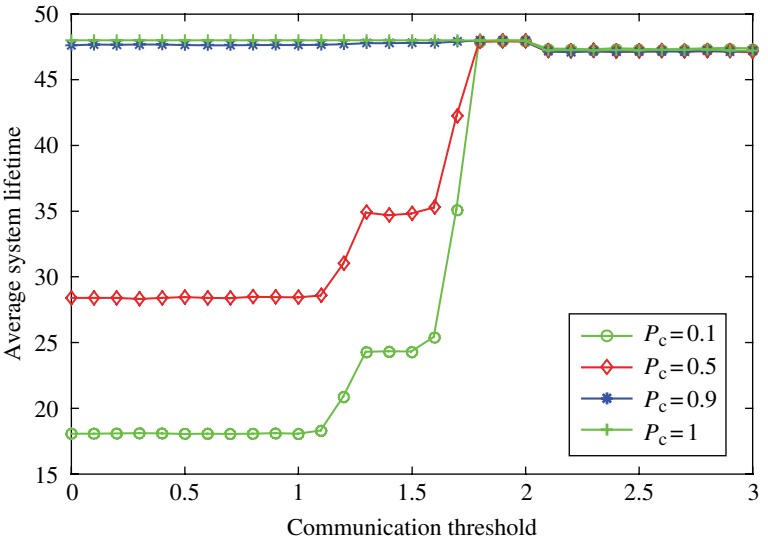


FIGURE 9.12 Effect of varying the threshold of communication on the average system lifetime of executing the MDP policy ($N=2$).

trade-off between the benefit of executing the coordinated sensing strategy and the energy cost of communication required to follow the global policy.

Figures 9.13, 9.14, and 9.15 show the expected system power outage, utility, and system lifetime respectively from coordinated sensing using five sensors. In this case, the energy cost of communication is more pronounced. When the cost is high $(p_c = 0.1)$, decreasing the amount of communication (by increasing the threshold) improves system performance (lower expected power outage) as executing the optimal coordinated sensing policy is less important than reducing energy expenditure of communication. When the communication cost is zero $(p_c = 1)$, increasing the threshold increases the expected power outage of the system since the coordinated sampling policy is increasingly approximated. At an intermediate communication cost $(p_c = 0.9)$, the expected power outage shows a minimum when the communication threshold is at approximately 3. This represents the best trade-off between the energy cost of communication and the benefit of communication in executing the coordinated sensing policy.

9.7.4 Effect of MDP Parameters on Policy

9.7.4.1 Effect of Penalty for Running Out of Power We changed the magnitude of the negative penalty that was used in the reward function (R_{powerout}) to study its impact on the policy. Figure 9.16 shows the system energy outage percentage for a range of penalties in the two-sensor case. The plot shows the benefit of (full) communication. Figure 9.17 shows the average utility of the corresponding policies.

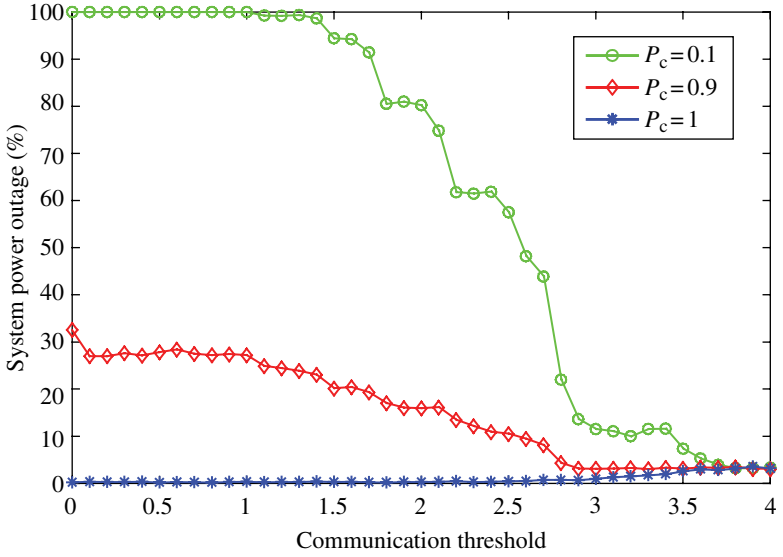


FIGURE 9.13 Effect of varying the threshold of communication on the system power outage ($N = 5$).

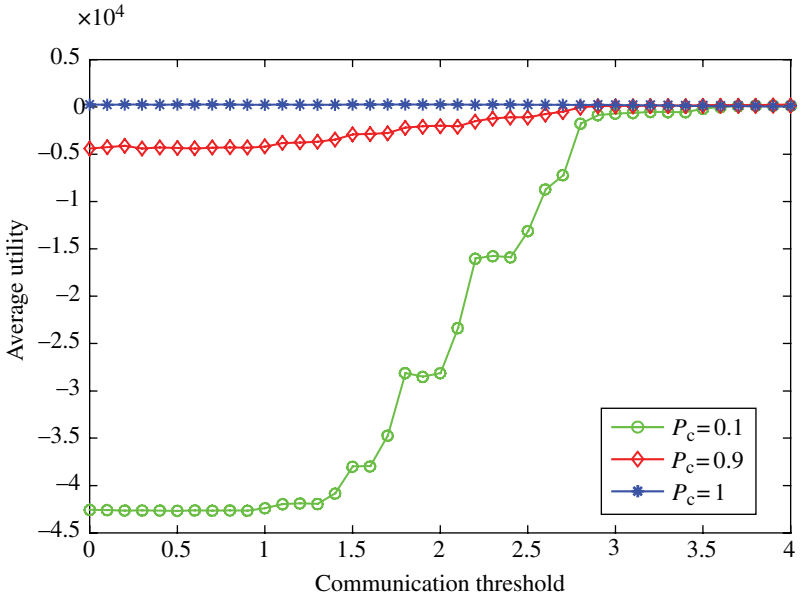


FIGURE 9.14 Effect of varying the threshold of communication on the average utility of executing the MDP policy ($N = 5$).

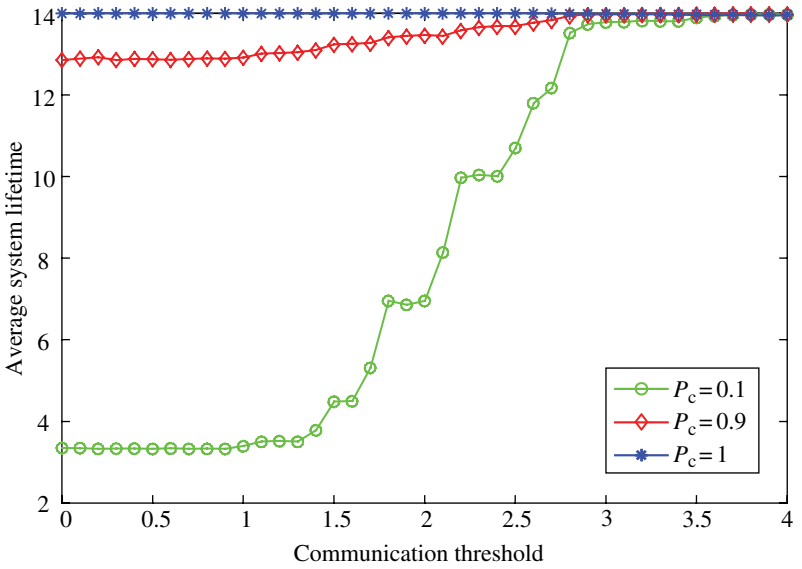


FIGURE 9.15 Effect of varying the threshold of communication on the average system lifetime of executing the MDP policy ($N = 5$).

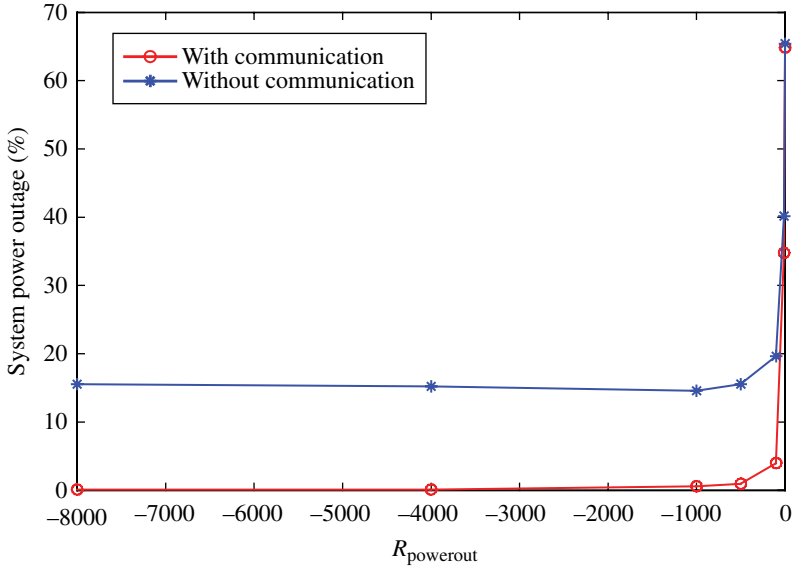


FIGURE 9.16 Effect of penalty in the reward function on the system power outage percentage for both the full- and no-communication case.

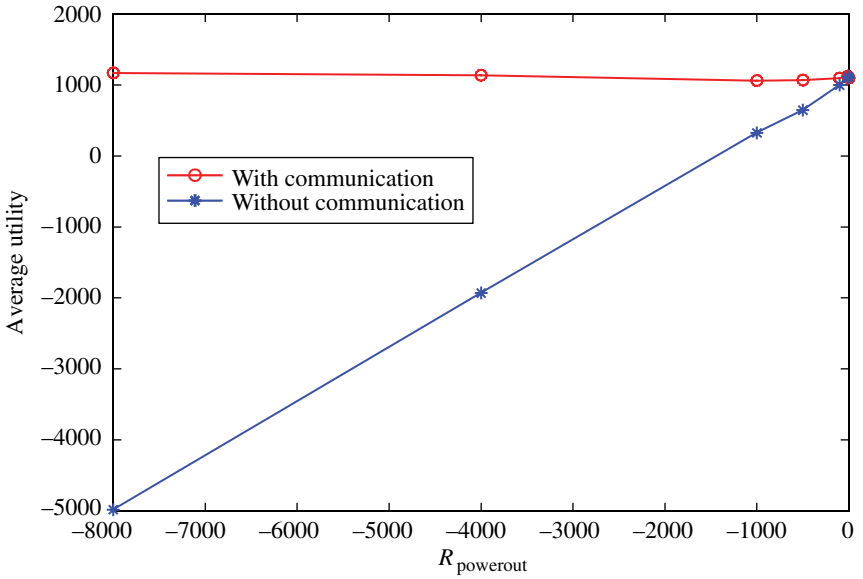


FIGURE 9.17 The utility of the MDP policy with different penalties in the reward function for both the full- and no-communication case.

The utility of full communication is higher than the no-communication case all the time. When the magnitude of the penalty becomes more negative, the difference of utility for full-communication case is more obvious.

9.7.4.2 Effect of Energy Consumption Rate on Lifetime In the MDP formulation, the rate at which energy is consumed at a sensor when sampling at a particular rate is modeled stochastically. We derived analytical equations to relate the energy consumption rate at a particular sampling rate to the expected system lifetime. In these derivations, we only consider the case where the sensors sample at a fixed rate. The goal is to understand the relationship between the stochastic energy consumption rate and the expected system lifetime in order to help us decide the appropriate discretization to be used while formulating the MDP model (i.e., to decide how finely time and energy have to be subdivided in the model).

Let p denote the probability that the energy level increases by one at a particular sampling rate. Let P denote the number of discrete energy levels and T the number of discrete time levels. We now derive the expected lifetime for such a system. The lifetime is the number of time-steps that at least one of the sensors has not run out of energy. Note that the system lifetime could be an arbitrarily large number with some small but nonzero probability. As the time model in the MDP is finite, we set the lifetime to T for all cases when the system survives beyond the maximum time level, T, in the model. This enables us to compare the analytical results with finite simulations.

Let $Pr_{\text{fail}}(t,p)$ denote the probability that the lifetime of a single sensor is less than t. p is the probability that the energy level of the sensor increases in one time-step. Then,

$$Pr_{\text{fail}}(t,p) = \sum_{j=P-1}^{t-2} \binom{j-1}{P-2} p^{P-1} (1-p)^{j-P+1} \tag{9.14}$$

Define $Pr_{\text{fail}}(t,p) = 0$ if $P-1 > t-2$. This is the probability that the energy level increased at least $(P-1)$ times in $(t-1)$ time-steps. The probability of the lifetime being exactly $(t-1)$ is

$$Pr_{\text{fail,t}}(t,p) = Pr_{\text{fail}}(t,p) - Pr_{\text{fail}}(t-1,p) \tag{9.15}$$

The expected lifetime of the single sensors is given by

$$\sum_{t=P+1}^{T} (t-1) \times Pr_{\text{fail},t}(t,p) + T \times \left(1 - \sum_{t=P+1}^{T} Pr_{\text{fail},t}(t,p)\right) \tag{9.16}$$

The second term denotes the case where the sensor survives for more than T time-steps. The probability of the system (two sensors) lifetime is $(t-1)$ is given by

$$Pr_{\text{fail,system}}(t,p) = 2 \times Pr_{\text{fail},t}(t,p) \times Pr_{\text{fail}}(t,p) - \left(Pr_{\text{fail},t}(t,p)\right)^2 \tag{9.17}$$

The expected system lifetime is then given by

$$\sum_{t=P+1}^{T} (t-1) \times Pr_{\text{fail,system}}(t,p) + T \times \left(1 - \sum_{t=P+1}^{T} Pr_{\text{fail,system}}(t,p)\right) \tag{9.18}$$

We plot the expected system lifetime with different values for energy consumption rate and maximum available energy in Figure 9.18. The maximum number of time-steps is 20. The plot enables us to set the range of probabilities that should be used to model the energy consumption rate at different sampling rates, and the relative values of maximum energy and time in the MDP model.

9.7.5 Sensitivity to Errors in Model Parameters

In the earlier simulations, it was assumed that the stochastic model used during offline policy computation exactly matched the stochastic energy consumption model of the physical system during execution. Here, we generate the optimal MDP policy by underestimating and overestimating the true rate of energy expenditure of the sensor system (and which is used during execution). The amount of over- or underestimation is the difference in probability of energy increase (p_E) between the model used for computing the MDP policy and the true model used during execution. Figure 9.19 shows the effect that the model error has on the expected performance of the resulting policy, and Figure 9.20 shows the decrease in energy reserves over time with varying amounts of model error $(N = 2)$ when the agents do not communicate at all.

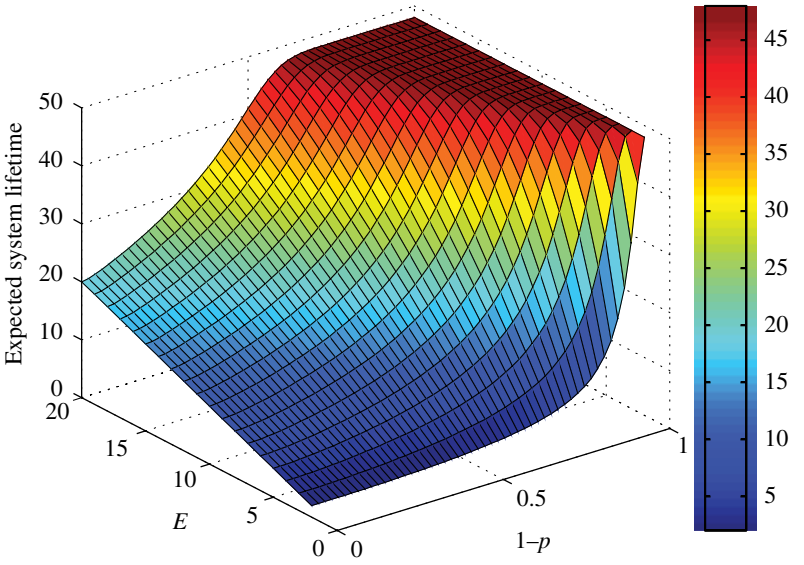


FIGURE 9.18 Expected system lifetime with different discretization of available energy (P) and energy consumption rate $\left(p_{\mathrm{E}} = 1 - p\right)$.

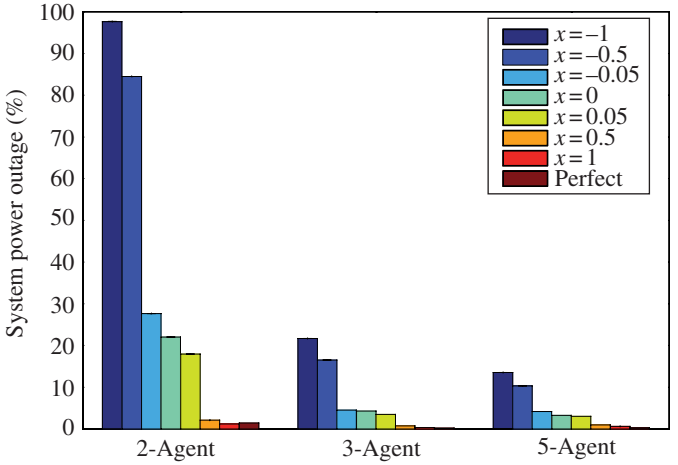


FIGURE 9.19 Probability of system power outage with varying amounts of model error $N = 2, 3, 5$.

(In these figures, an error of $x\,\%$ indicates that the probability of energy increase used for simulation is $1 + x$ times the probability used during policy calculation.) In these cases, the resulting policy is impacted by model errors. Figure 9.21 shows the corresponding decrease in energy reserves when the agents communicate during policy execution ($p_{\mathrm{c}} = 0.9$, threshold of communication = 1.75). These plots show

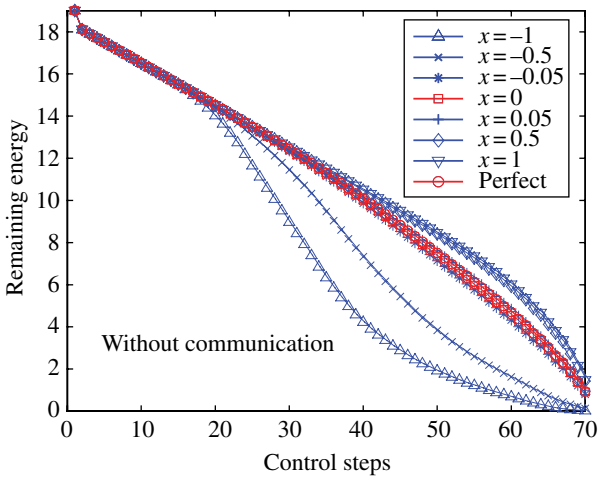


FIGURE 9.20 Remaining energy over time with varying amounts of model error and no communication ($N = 2$).

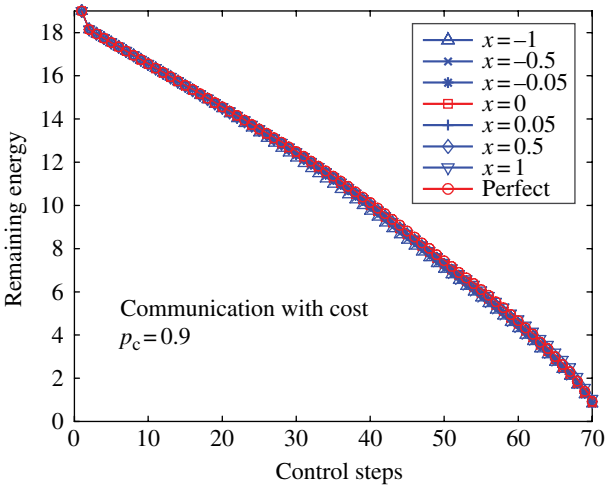


FIGURE 9.21 Remaining energy over time with varying amounts of model error and communication during policy execution ($N = 2$, $p_c = 0.9$, threshold of communication = 1.75).

that communication during execution offsets the inherent error in the model and removes the variation in the expected performance of the resulting policies. Figures 9.22 and 9.23 show the probability of the system running out of power and the expected lifetime when the model over or underestimates the energy expenditure rate. The ability to communicate local state information (local energy reserve) increases the system lifetime.

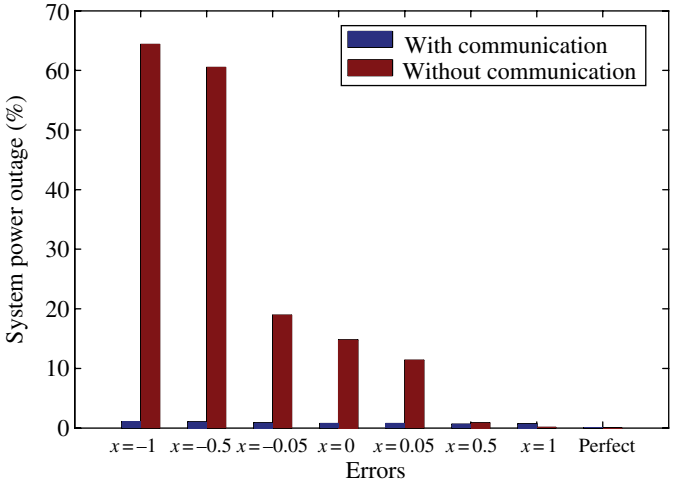


FIGURE 9.22 Probability of system power outage with varying amounts of model error ($N = 2$).

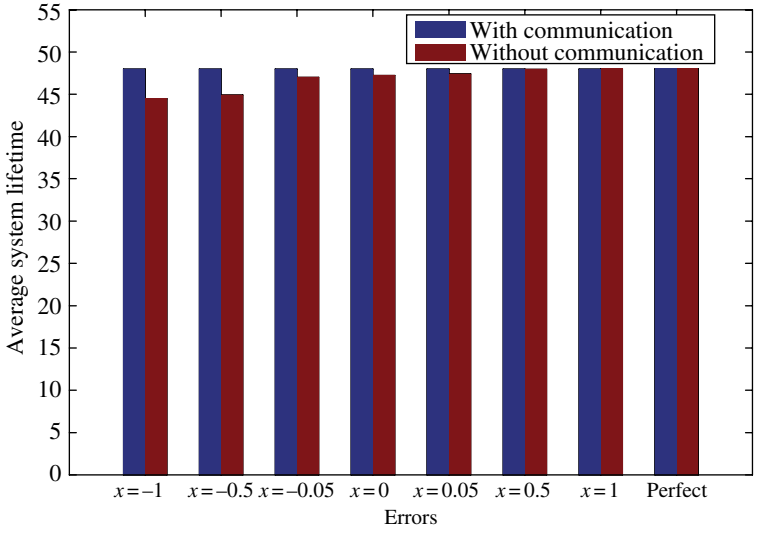


FIGURE 9.23 Average system lifetime with varying amounts of model error ($N = 2$).

9.7.6 Comparison with Random- and Fixed-Sampling Policies

We now compare the performance of the sensor sampling policies as computed by the MDP model with other fixed and random sampling policies. The fixed policies are as follows:

a. "Min": always sample at the lowest sampling rates.
b. "Max": always sample at the highest sampling rates.

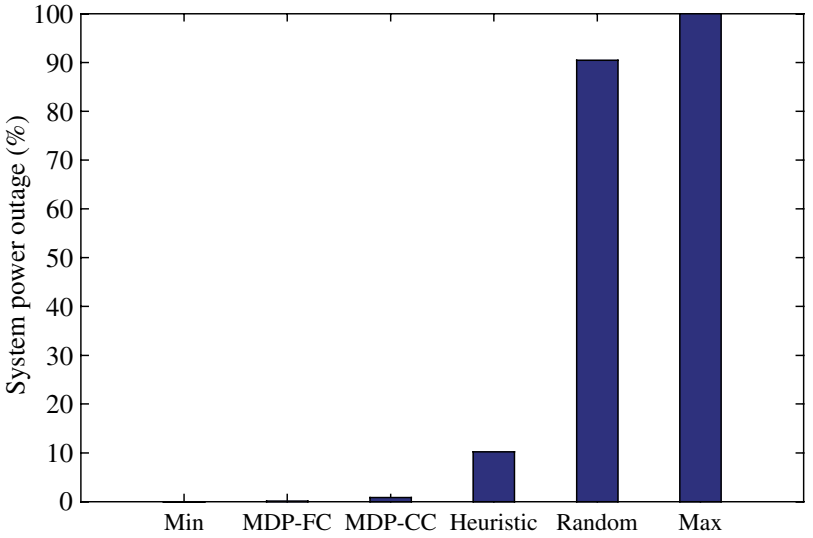


FIGURE 9.24 System power outage of MDP, Random, and Fixed policies ($N = 2$).

c. "Heuristic": sample at a rate determined by remaining energy (R_E), remaining time-steps (R_T), and energy consumption model ($p(a)$) as the following equation: $R_T = \dfrac{R_E}{1 - p(a)}$.

The random policy (Random) is to sample at a random rate at every time-step. The MDP policies are executed with communication with no energy cost (MDP-FC) and with a nonzero cost of communication (MDP-CC). The probability of system power outage is shown in Figure 9.24, the corresponding average system lifetime is shown in Figure 9.25, and the average utility for a system with two sensors is shown in Figure 9.26. Note that the MDP-FC policy results in a power outage probability as low as the minimum sampling rate case (the best possible) but while enabling the sensors to sample at higher sampling rates toward the end of the desired (the highest utility). The MDP policy is thus able to use the current time and the desired lifetime to increase the sensor sampling rates when it becomes likely that the system will not run of power before the desired duration. Figures 9.27 and 9.28 show the corresponding comparison of the expected system power outage and average system lifetime for a system with five sensors. Note that though the heuristic method has a low expected system power outage and long system lifetime, the average utility of the measurements (as computed using the rewards in the MDP state model) are lower than that of the MDP sampling method (Figs. 9.26 and 9.29). And the utilities for Random and Max policies are much lower than other policies. Hence in Figures 9.26 and 9.29, we only show the utilities for Min, MDP-FC, MDP-CC, and Heuristic policies.

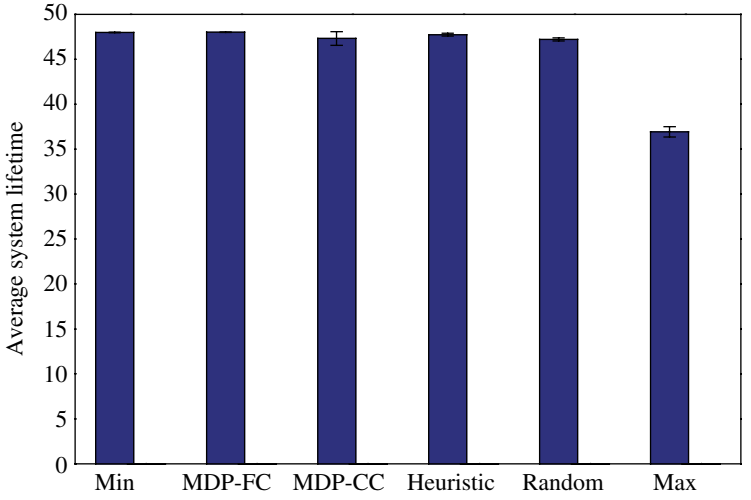


FIGURE 9.25 Average system lifetime of MDP, Random, and Fixed policies ($N = 2$).

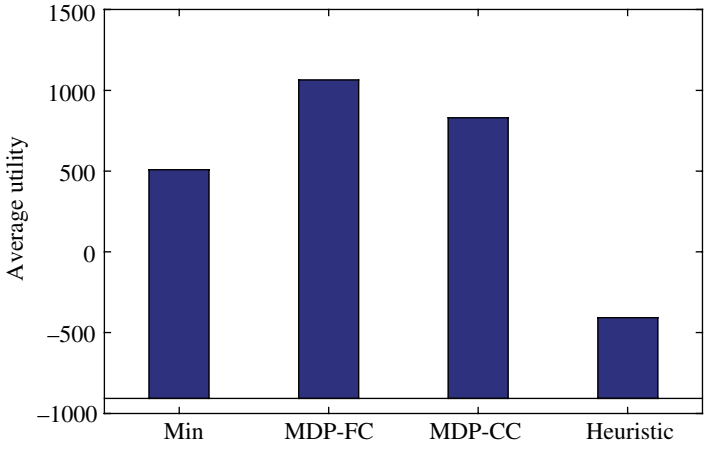


FIGURE 9.26 Average utility of sensor measurements as computed using the rewards in the MDP state model for different sampling strategies. The utilities of the Random and Max sampling policies are much lower and hence not shown ($N = 2$).

9.8 CONCLUSIONS

We have shown how the MDP framework can be used as the basis for coordinated sampling in a sensor network. Our approach enables an optimal policy to be computed before deployment under the assumptions of full observability of the local state of all sensors. We have formulated a communication scheme that enables individual sensors to execute this global policy by communication with other sensors only when the

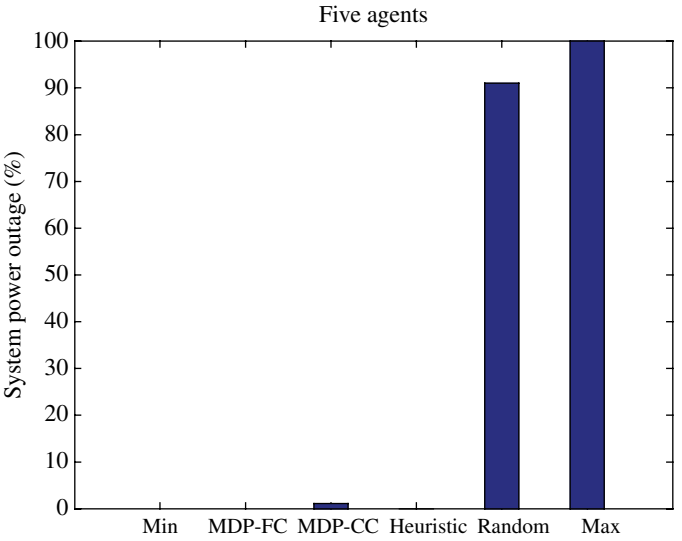


FIGURE 9.27 System power outage of MDP, Random, and Fixed policies ($N = 5$).

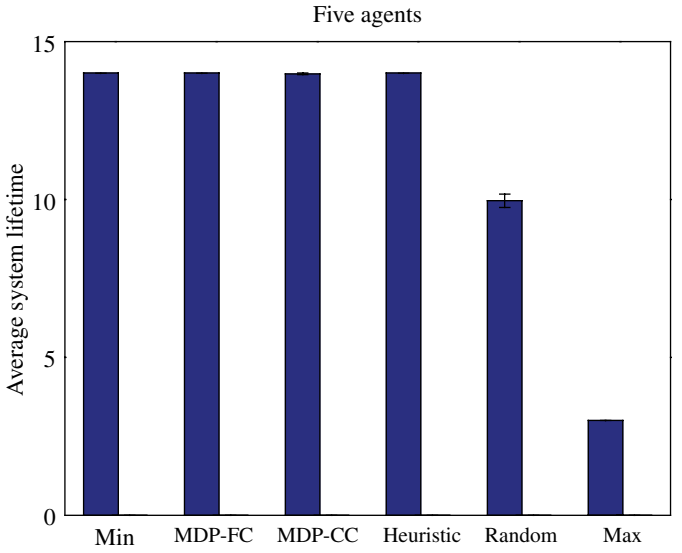


FIGURE 9.28 Average system lifetime of MDP, Random, and Fixed policies ($N = 5$).

expected value of information to be gained from the communication is high. We have shown simulation results that characterize the performance of this control framework. This method is suitable for networks of relatively few sensors and where the computational capabilities and energy reserves at each node are limited.

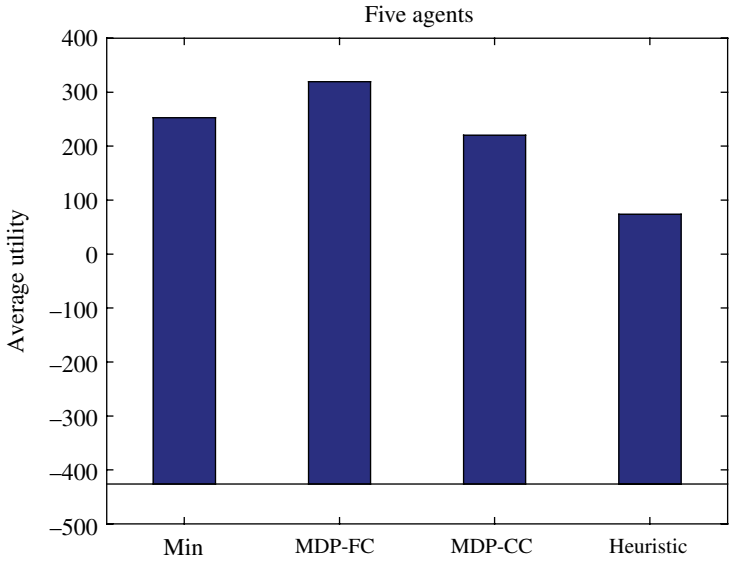


FIGURE 9.29 Average utility of sensor measurements as computed using the rewards in the MDP state model for different sampling strategies ($N = 5$).

ACKNOWLEDGMENT

This work was supported by NSF grant no. 0615132 from the Division of Computer and Network Systems.

REFERENCES

[1] C. S. Raghavendra, K. M. Sivalingam, and T. Znati, Eds., Wireless Sensor Networks. Berlin: Springer, 2006, pp. 3–20.

[2] G. Plasqui and K. R. Westerterp, "Physical Activity Assessment with Accelerometers: An Evaluation against Doubly Labeled Water," Obesity, vol. 15, pp. 2371–2379, 2007.

[3] M. Venugopal, K. E. Feuvrel, D. Mongin, S. Bambot, M. Faupel, A. Panangadan, A. Talukder, and R. Pidva, "Clinical Evaluation of a Novel Interstitial Fluid Sensor System for Remote Continuous Alcohol Monitoring," IEEE Sensors Journal, vol. 8, pp. 71–80, 2008.

[4] A. Talukder. Motes for Mobile Health Monitoring and Tele-medicine. Crossbow Solutions, vols. 1–4, 2005. Available: http://ceng.usc.edu/techreports/2009/Raghavendra%20CENG-2009-6.pdf (accessed February 22, 2016).

[5] A. Panangadan, S. M. Ali, and A. Talukder, Eds., Markov Decision Processes for Control of a Sensor Network-based Health Monitoring System, in *Proceedings of the Seventeenth Innovative Applications of Artificial Intelligence Conference*. Pittsburgh: AAAI Press, Menlo Park, CA, 2005.

[6] A. P. Chandrakasan, and R. W. Brodersen, Low Power CMOS Digital Design. Norwell, MA: Kluwer Academic Publishers, 1996.

[7] V. Raghunathan, P. Spanos, and M. Srivastava, "Adaptive power-fidelity in energy aware wireless embedded systems," in *IEEE Real Time Systems Symposium*, London, UK, December 2–6, 2001.

[8] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in *Proceedings of IEEE INFOCOM*, New York, June 23–27, 2002.

[9] R. Willett, A. Martin, and R. Nowak, "Backcasting: adaptive sampling for sensor networks," in *Proceedings of the 3rd ACM/IEEE Symposium on Information Processing in Sensor Networks (IPSN)*, Berkeley, CA, April 26–27, 2004.

[10] P. Aghera, T. S. Rosing, D. Fang, and K. Patrick, "Energy Management in Wireless Healthcare Systems," in *8th ACM/IEEE International Conference on Information Processing in Sensor Networks*, San Francisco, USA, 2009.

[11] A. Krause, M. Ihmig, E. Rankin, D. Leong, S. Gupta, D. Siewiorek, A. Smailagic, M. Deisher, and U. Sengupta, "Trading off Prediction Accuracy and Power Consumption for Context-Aware Wearable Computing," presented at the Ninth IEEE International Symposium on Wearable Computers (ISWC), Galway, Ireland, November 6–10, 2005.

[12] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram, "Markov-Optimal Sensing Policy for User State Estimation in Mobile Devices," presented at the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Stockholm, Sweden, 2010.

[13] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram, "The Tradeoff between Energy Efficiency and User State Estimation Accuracy in Mobile Sensing," in *The 1st Annual International Conference on Mobile Computing, Applications, and Services (MobiCASE)*, San Diego, USA, 2009.

[14] S. Zilberstein, R. Washington, D. S. Bernstein, and A.-I. Mouaddib, "Decision-Theoretic Control of Planetary Rovers," in Plan-Based Control of Robotic Agents. vol. 2466, M. Beetz, J. Hertzberg, M. Ghallab, M.E. Pollack (eds.). LNAI, New York, 2002, pp. 270–289.

[15] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman, "Solving Transition Independent Decentralized Markov Decision Processes," Journal of Artificial Intelligence Research, vol. 22, pp. 423–455, 2004.

[16] R. Emery-Montemerlo, G. Gordon, J. Schneider, and S. Thrun, "Approximate solutions for partially observable stochastic games with common payoffs," in *Proceedings of AAMAS*, Columbia University in New York City, July 19–23, 2004.

[17] D. Szer, F. Charpillet, and S. Zilberstein, "MMA*: A heuristic search algorithm for solving decentralized POMDPs," in *Proceedings of IJCAI*, Edinburgh, Scotland, August 1, 2005.

[18] D. S. Bernstein, S. Zilberstein, and N. Immerman, "The complexity of decentralized control of MDPs," in *Proceedings of UAI*, Stanford University, Stanford, CA, June 30–July 3, 2000.

[19] P. Xuan, V. Lesser, and S. Zilberstein, "Communication Decisions in Multi-agent Cooperation: Model and Experiments," in *Fifth International Conference on Autonomous Agents*, Montreal, Canada, 2001.

[20] S. Williamson, E. Gerding, and N. Jennings, "A principled information valuation for communications during multi-agent coordination," in *Proc AAMAS Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains*, Estoril, Portugal, 2008.

[21] C. V. Goldman and S. Zilberstein, "Optimizing Information Exchange in Cooperative Multi-agent Systems," in *Second International Joint Conference on Autonomous Agents and Multiagent Systems*, Melbourne, Australia, 2003, pp. 137–144.

[22] R. Nair, M. Roth, and M. Yohoo, "Communication for Improving Policy Computation in Distributed POMDPs," in *Third International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY, 2004, pp. 1098–1105.

[23] G.-Z. Yang and M. Yacoub, Body Sensor Networks: Institute of Biomedical Engineering Imperial College, London, 2006.

[24] D. Malan, T. Fulford-Jones, M. Welsh, and S. Moulton, "CodeBlue: An Ad Hoc Sensor Network Infrastructure for Emergency Medical Care," in *MobiSys Workshop on Applications of Mobile Embedded Systems (WAMES)*, Boston, MA, June 6–8, 2004.

[25] A. Talukder, A. Panangadan, L. Chandramouli, and S. Ali, "Optimal sensor scheduling and power management in sensor networks," in *Invited Talk at the SPIE Defense and Security Symposium, Optical Pattern Recognition XVI*, Orlando, FL, 2005.

[26] A. Joshi, P. Kowey, E. Prystowsky, D. Benditt, D. Cannom, C. Pratt, A. McNamara and R. Sangrigoli, "First experience with a Mobile Cardiac Outpatient Telemetry (MCOT) system for the diagnosis and management of cardiac arrhythmia", The American Journal of Cardiology, vol. 95, no. 7, pp. 878–881, 2005.

[27] L. Li and M. L. Littman, "Lazy approximation for solving continuous finite-horizon MDPs" in *Twentieth National Conference on Artificial Intelligence (AAAI)*, Pittsburgh, PA, 2005, pp. 1175–1180.

[28] C. Boutilier, R. Dearden, and M. Goldszmidt, "Exploiting structure in policy construction," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995, pp. 1104–1111.

[29] P. Liberatore, "On Polynomial Sized MDP Succinct Policies," Journal of Artificial Intelligence Research, vol. 21, pp. 551–577, 2004.

PART III

NANOSCALE WIRELESS COMPUTING IN MEDICINE

10

AN INTRODUCTION TO NANOMEDICINE

Amber Bhargava[1], Janet Cheung[1], Mary Mehrnoosh Eshaghian-Wilner[1], Wan Lee[2], Kodiak Ravicz[1], Mike Schlesinger[1], Yesha Shah[1], and Abhishek Uppal[1]*

[1]*EE-Systems, University of Southern California, Los Angeles, CA, USA*
[2]*Mechanical and Aerospace Engineering, California State University, Long Beach, CA, USA*

10.1 INTRODUCTION

Nanotechnology promises a host of innovative solutions in medicine through precise, targeted operations on the cellular level. For more than a decade, researchers have been searching for ways to realize these solutions and overcome the difficulties encountered when attempting to use nanotechnology for the treatment of various ailments. Recently, researchers have begun to look into nanorobotics, a relatively new subset of nanotechnology that involves the control of multifunctional nanosystems. Nanorobots could integrate the diagnosis and treatment of cancer into a cohesive, potentially noninvasive unit. Although the field of nanorobotics remains largely theoretical, a number of advances have been made in the past few years that lead us to believe that a multifunctional medicinal nanorobot could be possible. In order to achieve this goal, we need to know what has already been achieved and what still remains to be done.

Nanotechnology as a whole is a relatively recent development in scientific research. First defined by Norio Taniguchi from the Tokyo Science University in 1974, nanotechnology "mainly consists of the process of separation, consolidation, and deformation of materials by one atom or molecule [1]." Prior to the year 2000,

*Authors are listed alphabetically by last name, as opposed to by order of contribution.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

nanotechnology focused on passive nanostructures; now, active nanostructures and systems such as targeted drugs and nanorobots are being studied. Today, a widely accepted definition for nanotechnology is "engineering of functional systems at the molecular scale," according to the Center for Responsible Nanotechnology (CRN) [2]. The molecular scale or nanoscale refers to technology smaller than a 100 nm.

Nanomedicine is a subdiscipline of nanotechnology first defined in late 1999 and early 2000. Robert A. Freitas first comprehensively addressed the topic in his book titled *Nanomedicine*, which addresses the technical issues involved in the medical applications of molecular nanotechnology and medical nanodevices. Currently, there is no internationally agreed-upon definition for nanomedicine. For the purposes of this chapter, we apply the accepted definition of nanotechnology to medicine. We define nanomedicine as the use of nanoscale devices or materials for diagnosing and curing diseases by actively interacting at the molecular level within a cellular system. This definition is fairly broad, so we further restrict ourselves by including only those medicines whose active ingredients were specifically engineered with the intent of functioning on the nanoscale. This requirement rejects medications such as Chap Stick or sunscreen, which were not designed with a focus on the nanoscale. We believe this definition adheres most closely to the original intent of the term. Nanomedicine is a broad field that encompasses multiple topics from disease detection, drug delivery, and nanodevices to nanoimaging and clinical issues, all of which will be addressed in detail.

Nanomedicine is exciting because many of its findings—though often based on preliminary or even theoretical experimentation—promise results that are challenging at the macro scale. For example, many devices and materials being studied experimentally may be used to detect faulty cells via antibody conjugation. Carbon nanotubes (CNTs), for instance, are capable of detecting particular deoxyribonucleic acid (DNA) mutation sequences that could give rise to cancer. Another viable application of nanoparticles in medicine is molecular imaging. Specific biomolecules can be tagged and quantitatively analyzed via nanoparticles that glow under infrared (IR) light. Both of these approaches have exciting and novel applications in medicine, but they only scratch the surface of how nanotechnology can potentially influence the future of medicine. In order to imagine the true scope, we can look at Nanorobotics.

Nanorobotics is a multidisciplinary field that requires knowledge of physics, chemistry, biology, computer science, and electrical engineering. Nanorobots must possess the capabilities of actuation, control, communication, and interfacing between the organic, inorganic, and biotic realms [3]. There are many directions from which to approach the problem of creating a viable nanorobot. Some researchers focus on bionanorobotic systems made of biological components such as proteins and DNA, while others focus on inorganic materials like complementary metal oxide semiconductor (CMOS) and CNTs at the cellular level. Still others choose to focus on hybrid nanosystems, combining the organic and inorganic domains in order to obtain the best of both domains into one nanorobot.

First, this chapter will address the types of diseases investigated and the nanotechnology involved in medical treatment. Next, it will discuss the current detection and treatment methods used by researchers. Then, the chapter will conclude with current

nanotechnology used in cancer treatment, issues with biocompatibility, and current research. Specific examples of computer modeling will be discussed in depth.

10.2 NANOMEDICAL TECHNOLOGY

In this section, we will discuss nanotechnology as it applies to medicine, including different types of nanorobots and some of the challenges inherent to building robots on the nanoscale.

Nanorobotics is an emerging technological field that involves the development of robots whose components are at or close to the nanometer scale [4]. Nanorobots designed to operate inside the human body should be less than 800 μm in diameter in order to traverse the bloodstream [5]. So far, no complex nanorobot has been fabricated—current medical robots are constrained to the micrometer level. Thus, the construction of robots on the nanoscale is merely a theoretical model. In this section, we present a short overview of medical nanotechnology, including nanorobots, nanoparticles, and CNTs as nanodevices related to medicine. We also take a look at microscale devices that represent the current state of the art in medical implantable and noninvasive technology (Fig. 10.1).

Treatments based on nanorobots are expected to have two major advantages over traditional approaches: concentration and precision. These advantages enable therapies to become more efficient and cause fewer side effects because of their ability to concentrate drugs on a single tissue area or cell. In addition to these advantages, nanorobots also have the ability to respond collectively to environmental stimuli in order to accomplish their designated tasks.

There are two main approaches to developing an assembled nanorobotic system: organic and inorganic. Some organic nanorobots are based on adenosine triphosphate (ATP) and DNA molecular assembly and function. Another subset of organic nanorobots is nanorobots based on bacteria, which have been proposed experimentally for drug delivery. Inorganic nanorobots are essentially nano-electromechanical systems (NEMS). As technology has developed, some approaches have begun to combine both organic and inorganic elements, creating a more advanced robotic system known as a hybrid nanosystem.

We believe that inorganic nanorobots are well-suited as candidates for nanomedicine, and so we will concentrate our survey on inorganic nanorobots. There are several differences between inorganic and organic nanorobots, which determine how

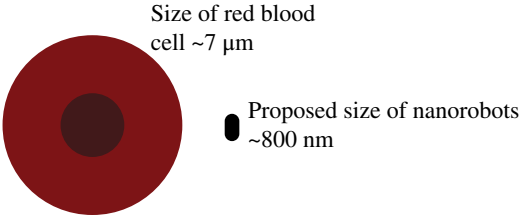


FIGURE 10.1 Size of nanorobots relative to red blood cells.

these two subsets will approach the many problems inherent to medical nanorobots. For example, inorganic nanorobots could locate the exact position of a cell using the intensified magnetic property from the force of attraction inside the human body. In other words, they can isolate a target cell from a normal cell and at the same time allow for targeted drug delivery via a nanorobot at that destination. An organic nanorobot may be able to perform the same task using DNA biomarkers. However, the inherent programmability of inorganic nanorobots based on CMOS technology and the ability to leverage the existing control structures of macro-sized robots make inorganic nanorobots the most likely to succeed in performing the complex, precise tasks required of medical nanorobots.

Still, there are many challenges to building a multifunctional inorganic nanorobot. In order to create a robot in the nanoscale, we must take into account sensing, actuation, control, data transmission, power, and interfacing across spatial scales as well as between the organic/inorganic and biotic/abiotic realms [6].

Although inorganic robots do not exist in the nanoscale today, we can look at microscale and larger medical robots that can be used as test cases for extending functionality into the nanoscale. Additionally, research on surgically implanted devices can provide information on how nanorobots may be able to solve the problems of biocompatibility, energy transfer, and data transfer.

One such area is diagnostic medicine in the digestive tract. Medical device companies have used capsule endoscopes for viewing and diagnosing diseases. These devices have onboard cameras and light-emitting diodes (LEDs), and can send video wirelessly out of the body [7]. However, they are still limited in scope as they pass through the body without self-propulsion, drawn solely through peristaltic pressure, and therefore cannot complete tasks such as truly targeted drug delivery or a thorough search for cancerous regions. There are groups searching for ways to overcome these challenges, however, by developing motors [7], controlling several microrobots in a system [8], and looking at the potential for microrobots that could dispense medicine into the stomach for months at a time [9].

Although microscale and nanoscale robots face very different problems in terms of motion and energy supply [10], microscale medical devices are important to the emerging field of nanorobotics both as a technological precursor and as a testing ground for the human side of medical nanorobotics. Microrobots open the discussion of the legal issues of data security and how the medical and device-making community will deal with implantable devices that are constantly streaming important personal data.

Today, nanorobotics is rapidly becoming a reality. Penn State published an article in February 2014, showing nanomotors individually controlled inside living cells [11]. This is an important step, as it addresses both control methods and biocompatibility. However, there are still many steps to be taken before nanorobots are viable inside the human body.

Current nanorobot components, control methods, and developing manufacturing methods provide platforms and possibilities for building multifunctional nanorobots that may be able to swim through vessels, detect and destroy cancer

cells, or send pictures back to their controlling devices with accuracy, controllability, and precision. As previously stated, we wish to develop nanorobots in order to perform precise drug delivery.

10.3 DETECTION

In Section 10.2 we looked at nanotechnology as it applies to medicine, different types of nanorobots, and why we chose to focus this chapter on inorganic nanorobots. In Section 10.3, we will discuss a number of different nanostructures we can use to detect the presence of cancer or other diseases in the body.

Nanomedicine can assist in the treatment of a wide range of diseases. Our research focuses on those that can be treated by nanorobots in the bloodstream, including diabetes, cardiovascular disease, neurodegenerative disease, and cancer, among others. While a major part of nanomedical research is still in its infant stages, significant research and experiments have been performed on these and other diseases. While all of these diseases are a threat to public health and could potentially be treated or cured by nanomedicine, our research focuses specifically on cancer.

Cancer is a group of diseases characterized by uncontrolled cell growth. The body does not regulate cancer cells as it does healthy ones, thereby allowing them to replicate at a rapid rate without being killed off as they normally would. Cancer cells replicate so uncontrollably due to damaged DNA, which prevents the cell's natural death and causes changes in material (physical, electrical, or chemical) properties within the cell.

Early detection and prevention are often the most effective cures for many diseases, such as cancer. Currently, early detection is accomplished by identifying biomarkers, an indicator of a biological state of disease. Biomarkers can be DNA, RNA, or a protein and its fragments. In the future, even earlier detection may be possible thanks to ubiquitous computing and constant monitoring of at-risk patients. Today's detection tools, however, function in response to changes in material properties, such as those caused by the damaged DNA in cancer cells. Detection tools under discussion in this chapter are nanowires, CNTs, nanoscale cantilevers, various nanoparticles (gold and magnetic, among others), quantum dots, and nanorobots. In this section, detection tools for biomarkers are discussed.

10.3.1 Nanowires

Nanowires inherently have excellent selectivity and specificity properties. For example, nanowires can be laid down across a microfluidic channel, not unlike a filter or a spider's web. As cells or particles flow through it, nanowires can sense and pick up tumor cells. Nanowires can be coated with a probe such as an antibody or oligonucleotide—a short stretch of DNA that can be used to recognize specific DNA/RNA sequences—that binds to a target protein on the enemy cell. Proteins that bind to the antibody will change the nanowire's electrical conductance, which can be measured by a detector [12].

10.3.2 Carbon Nanotubes

CNTs are also being used as DNA biosensors. They scan down DNA and look for single polymorphic nucleotides that can lead to a possible detection of an individual who may develop diseases in the future. This application uses self-assembled CNTs and searches for covalently bonded DNA oligonucleotides. When hybridization between the probe and the target DNA sequence occurs, the voltammetric peak picks up the change [13]. DNA biosensors being developed for future use are more efficient and more selective than current detection methods.

10.3.3 Nanoscale Cantilevers

Another potential tool is the nanoscale cantilever. Similar in structure to rows of diving boards, these cantilevers are constructed using semiconductor lithographic techniques [14] and coated with antibodies that specifically target molecules produced by cancer cells. These target molecules bind to the antibodies on the cantilever, causing a change in its physical properties. For quantitative analysis, researchers can study the binding in real time. These cantilevers are exceptionally sensitive and can detect single molecules of DNA or protein, hence providing precise detection methods for cancer-related molecules. Figure 10.2 shows how nanoscale cantilevers detect proteins produced by cancer cells.

10.3.4 Gold Nanoparticles

Gold nanoparticles (GNPs) have been emerging as powerful imaging labels and contrast agents—hence effective detectives. GNPs readily get conjugated to antibodies and other proteins due to the affinity of the functional group (—SH) for their gold surface. They are especially effective when targeting cancer cells [15]. GNPs are used for photothermal therapy, where tunable optical properties cause them to convert laser light into heat and destroy cancer cells [16, 17]. In addition to spherical GNPs, gold nanoshells and gold nanorods have been applied to biomarker detection [18–20]. They can absorb and emit light at near-infrared (NIR) region, providing deep tissue penetration.

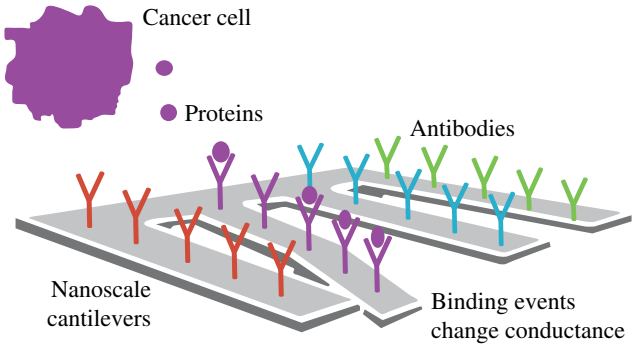


FIGURE 10.2 Detection of cancer cells by nanoscale cantilevers.

10.3.5 Magnetic Nanoparticles and Magneto-Electric Nanoparticles

Magnetic nanoparticles have been widely used as an imaging tool. The concentration of cancer biomarkers can be measured by taking advantage of the the aggregation of magnetic nanoparticles that occurs when target molecules are present and that decreases transverse relaxation time. This allows for in vivo, local monitoring for cancer biomarkers and possible continuous monitoring. Magneto-electric nanoparticles have also proven to be a promising new technology. In one specific study, researchers have been able to isolate a cancerous cell based on differences in the electric properties of its membrane [21]. This method of detection has been proven effective as applied to ovarian cancer cells, but may potentially be applied to other types of cancer as well. More on this topic can be found in Chapter 11.

10.3.6 Quantum Dots

Quantum dots can be linked to antibodies and combined to create arrays that are capable of detecting multiple substances simultaneously. They can be used to measure levels of cancer markers such as breast cancer marker Her-2, actin, microfibril proteins, and nuclear antigens [14]. Quantum dots are robust and very stable light emitters. The photochemical stability and the ability to tune broad wavelengths make quantum dots extremely useful for biolabeling [22].

10.3.7 The Medical Nanorobot

The ultimate tool of nanomedicine is the medical nanorobot—a robot of the size of a bacterium comprising multiple mechanical parts [23]. Although still a hypothetical concept, the medical nanorobot is a promising tool for the future of the field of medicine—a future in which artificially intelligent nanorobots can be fabricated to repair tissues, clean blood vessels and airways, and transform physiological capabilities. A more in-depth discussion of the technology behind nanorobots can be found in Section 10.2 as well as in Chapter 2. Section 10.4 will deal with actually treating the disease cells once they are detected.

10.4 TREATMENT

In Section 10.3, we identified six methods of cancer detection. In Section 10.4, we will discuss the problems with current cancer treatments, as well as methods that can be used to treat disease at the cellular level within the body.

Modern cancer treatment is anything but ideal. Conventional treatments such as chemotherapy, radiation, surgery, and immunotherapy destroy malignant tissue, but also damage benign tissue. These methods are found to be useful in remission, but depend on the concentration and delivery of the drug. Highly toxic drug concentrations that destroy the tumor cells can potentially kill the patient. Thus, the aggressiveness of chemotherapy treatment is usually determined by the dosage that the patient can withstand, rather than the dosage needed to eliminate all cancerous cells.

A more efficient approach to cancer treatment would be the destruction of cancer cells with little to no side effects on healthy cells. With the current trend of rapid developments in nanomedicine, it seems that this technology can be used for detection, analysis, and destruction of cancer cells more effectively and with more precision than is possible with current treatments.

As stated, the key problems of conventional technology are the method of drug delivery and the concentration of the drug cocktail required to destroy the cancerous cells. Nanotechnology allows drugs to be directed to the exact location where cancerous cells have been observed, thus killing malignant tissue while leaving healthy tissue relatively unaffected. This capacity for targeted drug delivery also allows for a significantly reduced concentration of the drug and a resulting reduction in the number of toxins in the body.

10.4.1 Nanoparticle Heating Method

The nanoparticle heating method for cancer therapy has great potential for treating selective cancer tissue. In this method, the nanoparticles conjugated with antibodies are injected into the human body and are allowed to position themselves around the cancer tissue. These particles are then heated by an external noninvasive heating source, destroying the targeted cancer tissues via thermal necrosis [24]. Capacitively coupled radio-frequency (RF) field sources [25] and NIR light sources [26] are two of the heating sources that could be used to heat up these nanoparticles. A major advantage of this technique is its lack of dependence on traditional drugs. As invading bacteria become less and less sensitive to antibiotics and other drugs, a new solution may soon be required to treat these bacterial infections. Nanoparticle heating may well be one solution. Two treatments that employ the nanoparticle heating method by way of IR and laser light are discussed in the following text.

10.4.1.1 Gold Nanoparticles As discussed in Section 10.3.4, one of the most researched nanomedical applications is GNPs. These nanoparticles bind to certain proteins that only cancer cells produce. GNPs are also sometimes referred to as "the drugs that deliver themselves" because of their frequent use in thermal necrosis. The downside of this method is that unless properly coated, the immune system would attack such particles because the body would treat them as unwanted foreign agents. See Section 10.5.1 on biocompatibility for more information on immune reactions and immunosuppression.

10.4.1.2 Metal Nanoshells Another novel approach uses metal nanoshells. These shells have the ability to capture and absorb light and are coated with a bioactive substance that binds them to cancer cells [27–29]. NIR light is used to heat up these shells, leading to the destruction of the cancer cells with minimal damage to adjacent healthy cells. The advantages of high sensitivity, specificity, and cost-effectiveness have made metallic nanoshells a particularly attractive choice for

modern-day cancer treatment. Alam and Massoud have developed an accurate analytical closed-form model for the frequency resonance and scattering characteristics of a single nanoshell [16].

10.4.2 Supermagnetic Beads

Apart from GNPs and metal nanoshells, another method that emphasizes the destruction of targeted cancer cells uses mono-sized supermagnetic beads. The beads are macroporous particles having narrow pores that contain magnetic materials (Fe_2O_3 and Fe_3O_4) distributed throughout their entire volume. In studies, two different types of beads having different magnetic mass susceptibility and different diameters were used: Dynabeads Pan Mouse IgG (diameter $4.5 \pm 0.2\,\mu m$, magnetic susceptibility $(16 \pm 3) \times 10^{-5}\,m^3/kg$) and Dynabeads Protein G (diameter $2.8 \pm 0.2\,\mu m$, magnetic mass susceptibility $(10 \pm 2.5) \times 10^{-5}\,m^3/kg$). These magnetizable beads aggregate under instantaneous pulsed magnetic forces and penetrate forcefully to effectively destroy the cancerous cells [21].

10.4.3 Magnetic Manipulation

Magnetic manipulation is another viable method to efficiently deliver medication. Recently, scientists have successfully used a direct current magnetic field to manipulate the membranes of magneto-electric particles. They then used this ability to load the same magneto-electric particles with paclitaxel, a common cancer drug, and deliver it to a mix of healthy and cancerous cells in vitro. The drugs, when released by the particles, passed through the membranes of the cancer cells and killed only the targeted cells. The healthy cells were left alive [30]. In the same study as previously mentioned in Section 10.3.5 of this chapter, magneto-electric nanoparticles were used to manipulate the magnetic field of the membrane of the tumor cell to allow large quantities of medication to pass through into the cell. The diseased cell was killed with no harm at all to surrounding healthy cells [21], as can be seen in more detail in Chapter 11. These results show early promise in specific targeting of cancer cells within the human body, and certainly deserve further research.

10.4.4 *E*-Cadherin

Nanorobots can also be used to analyze levels of *E*-cadherin in the body and target cancer cells based on the varying *E*-cadherin levels from one cell to the next [31, 32]. *E*-Cadherin is a calcium-regulated adhesion expressed in most normal epithelial tissues. It is associated with gland formation, stratification, and epithelial polarization. Perturbation or selective loss of this *E*-cadherin function results in the loss of intercellular adhesion, with possible consequent cellular transformation and tumor generation. The efficient use of nanorobots by proper control methodologies is one

of the methods under study for the treatment of cancer cells, as it is quite capable of differentiating normal cells from malignant cells by checking surface antigens. This in turn greatly reduces the probability of destroying normal cells.

10.4.5 Treatments for Other Diseases

While this chapter does focus primarily on cancer treatment, there are many other nanoscale techniques for treating diseases other than cancer. The following subsections will describe three of them briefly.

10.4.5.1 Transfection As opposed to the methods of drug delivery discussed throughout the rest of this section, transfection is a method of gene delivery. Transfection can be accomplished by both chemical and nonchemical processes. Chemical-based transfection can be divided into several categories, including cyclodextrin, polymers, liposomes, and nanoparticles. There are various nonchemical processes as well, including electroporation, sonoporation, and impalefection. Impalefection is different from these other nonchemical methods in that DNA is introduced into the cell through the use of nanomaterials. In this process, nanowires and similar nanoparticles are used as transport systems. Vertical arrays of nanowires are prepared by photolithography and plasma-enhanced chemical vapor deposition before they are coated in DNA containing the sequence meant for delivery. Target cells are cultured on these nanowire arrays, which proceed to impale the target cells as they settle on the surface of the nanowires. This method is able to deliver DNA directly to the cell cytoplasm of the nucleus of the cell, which then begins to act in accordance with the newly introduced sequence of DNA. Appropriately, the term "transfection" is a combination of both "impalement" and "infection" [33–36].

10.4.5.2 Tissue Engineering Nanomedical technology is being investigated as a possible method to further the field of tissue engineering. Nanotechnology could potentially be used to help repair damaged tissue through the use of suitable nanomaterial-based scaffolds and growth factors. If successful, tissue engineering may be able to replace conventional treatments such as organ transplants and artificial implants. Nanoparticles such as graphene, CNTs, molybdenum disulfide, and tungsten disulfide are being investigated as potential reinforcing agents to make strong, biodegradable polymeric nanocomposites for bone tissue engineering applications. As a result of the addition of these nanoparticles into the polymer matrix at low concentrations of about 0.2% by weight, the compressive and flexural mechanical properties of polymeric nanocomposites are significantly increased. These nanocomposites could potentially be used to further nanonephrology (defined as nanomedicine of the kidney), create strong, lightweight composite bone implants, or even weld arteries during surgery [37, 38].

10.4.5.3 Monitoring Medical monitoring is the practice of observing patients to either make sure they remain healthy or gathering data to assist in diagnosis or treatment. Traditionally, this has been done primarily in controlled environments where a doctor can physically observe the patient. Using current technology, however,

doctors can observe patients from anywhere by way of a small chip either implanted under the patent's skin, swallowed in pill form, or introduced to the body by other means. This is a major improvement over in-person monitoring, due to the fact that it enables constant monitoring. Constant monitoring of the patient allows doctors to gather more data from the patient, all while patients go about their normal daily routine. Both of these benefits can result in faster, more accurate diagnosis and treatment of disease. Unfortunately, due to the relatively large size of these modern implants as compared to objects on the nanoscale, they can only be implemented in limited locations throughout the body. This restricts their access to certain data that can only be obtained in locations in which they cannot be utilized. Nanomedical technology has the potential to overcome this obstacle by producing nanoscale implants that can be utilized virtually anywhere within the body. Using the concept of a nanoscale lab-on-chip implant, scientists can potentially gather data from tests ranging from continuous evaluation of blood sugar levels and neuroimaging to predicting both hypo- and hyperglycemic states as well as seizures. In vivo biomedical microelectromechanical system (Bio-MEMS)-based biosensors and serotonin biosensors may also be used in preventing an early detection of mental health disorders, such as depression [39, 40].

This section discussed just some of the multitude of treatment options currently under investigation. In Section 10.5, we will explore challenges introduced when introducing any of these foreign agents into the human body—a topic known as biocompatibility.

10.5 BIOCOMPATIBILITY

In earlier sections, we have covered methods to detect and cure cancer on the nanoscale. Achieving a treatment, however, is more difficult than that. The nanorobots need to be able to survive within the body for long enough to arrive at the cancerous cells and deliver their medication. In Section 10.5 we discuss the challenges presented by this requirement, commonly known as biocompatibility.

Biocompatibility, or "the ability of a material to perform with an appropriate host response in a specific situation," [41] is an important factor in the development of nanomedical robots. Any nanorobot designed to enter the human body must be biocompatible in order to function correctly, or its introduction will trigger unwanted immune responses in the body. Biocompatibility has traditionally been an important consideration in prosthetics and organ transplants, as a rejected transplant can have serious complications. Biocompatibility in nanomedicine, however, in part due to its large surface-area-to-volume ratio, introduces challenges unique to its scale [42]. This section will address both possible effects of nanorobots on the body and factors that determine biocompatibility on the nanoscale.

10.5.1 Immunostimulation

If medical nanorobots are not appropriately biocompatible, there is a risk of a number of complications within the host. The two very broad categories of immune responses to nanoparticles are immunostimulation and immunosuppression. If the nanorobots

trigger an immune response within the host, the reaction can be considered one of immunostimulation [43]. It is important to note here that biocompatibility requires an immune reaction appropriate to the desired function, not simply a non-reaction. Immunostimulation is not necessarily a negative reaction; it can, in fact, be the desired effect. For example, nanoparticles can be injected into the blood stream to aid the immune system in fighting off a disease [42, 43]. In fact, this method has been used since the advent of vaccines with non-nanoscale materials known as immunologic adjuvants [44, 45]. Immunostimulation can, however, be harmful to both host and nanorobots. Immunostimulation produces side effects including hypersensitivity to allergens, inflammation, fever, and other flu-like symptoms [43]. It can also aggravate other conditions, such as autoimmunity. Immunostimulation also results in the increased production of macrophages, which can attack and destroy nanorobots before they are able to accomplish their task within the body.

10.5.2 Immunosuppression

Introducing nanoparticles into the body does not always cause immunostimulation, but it can also trigger immunosuppression. As its name suggests, immunosuppression is the opposite of immunostimulation: it is defined as "suppression of the body's immune system and its ability to fight infections and other diseases [46]." Immunosuppression is vital to successful organ transplants and skin grafts, as well as autoimmunity treatments. Immunosuppression can make it easier for drug-delivery nanorobots to accomplish their jobs in the body, but also has the potential to harm or even kill the host [43]. Immunosuppression leaves the body open to attack, and is the symptom of diseases like HIV and AIDS that eventually lead to death. When nanorobots are used as drug delivery devices, as is the focus of this chapter, they will likely be most effective if they are seen as native entities with no immunostimulation or immunosuppression [44]. This is the goal of nanoscale biocompatibility.

10.5.3 Factors of Nanobiocompatibility

In the previous subsections, we discussed two major effects that the introduction of foreign nanorobots can have on the immune system: the immunostimulation and immunosuppression reactions. In this subsection, we present many of the factors that can influence these reactions. Biocompatibility on the nanoscale presents challenges not faced in traditional prosthetics or transplants. On the nanoscale, biocompatibility is determined by a number of factors such as size, amount, density, location, duration, geometric surface planes, and material, among others [34–36]. Material may be the largest factor in determining biocompatibility. Some materials, such as metal traces in certain CNTs [42], are known to cause undesired immune reactions, while others, such as diamond [47, 48] and possibly graphene [43], have been shown to have minimal biocompatibility issues in animals. Size is a key factor as well—even if a material is considered biocompatible on a larger scale, nanoparticles of the same material can potentially prove to be carcinogenic [48]. Also, the smaller the particle, the longer it tends to survive in the blood stream. According to Bastús et al., this is an

observable, universal trend [44]. Surface orientation plays a large role in determining biocompatibility as well, as each geometric plane presents a different physical structure to the surrounding environment. Immune responses differ between planes, causing biocompatibility to vary even within nanoparticles of the same material [47]. Biocompatibility also differs based on location. For example, nanorobots will encounter very different biological conditions if injected directly into the bloodstream, as opposed to being ingested orally or inhaled. Each set of conditions will affect the nanorobot differently, so the same nanorobot may be biocompatible in the bloodstream, but not in the stomach. Duration is also a major—if often overlooked—factor in biocompatibility. Nanorobots must have a pre-planned way to leave the body once they accomplish their task. They may be excreted normally, mimic pathogens so as to be removed by the immune system, or use other exit techniques [43]. If there is no exit strategy, nanorobots can build up in vital internal organs, such as the liver or spleen, causing side effects including liver failure and cancer [42–44].

Nanomedicine does have the potential to revolutionize the way medicine is practiced around the world, but it is clear that biocompatibility on the nanoscale is one of many major challenges that must be overcome before nanomedicine can be safely implemented on a large scale.

10.6 POWER

In Section 10.5.3, we discussed biocompatibility and the difficulties presented by this concept. But even if a nanorobot can detect cancer, treat cancer, and be fully biocompatible, it still needs a way to power itself. In this section, we will discuss recent progress in wireless power and how this idea can be directly applied to nanomedicine.

Unless the nanorobots of the future are designed to be totally passive—relying only on outside sources for movement and actions—they will need a source of power. Powering a nanorobot with wires as it circulates through the human bloodstream is obviously impractical, and nanoscale batteries would likely hold only a very limited charge and do not currently exist in forms useful in nanomedical applications. Ignoring external wires and internal batteries leaves us with a very likely solution to nanorobot power: wireless inductive power.

When powering a device like a nanorobot remotely, a four-part system consisting of a transmitter, rectifier, bandgap reference circuit, and matching network is required. The transmitter transmits AC power, which is converted to DC power by the rectifier. The bandgap reference circuit and a regulator use analog circuitry techniques to transform this DC voltage into a stable, non-temperature dependent voltage source usable by the device. The network must be impedance matched for the particular operating to minimize power loss, thereby providing more power to the chip [41] (Fig. 10.3). There are many known designs for wireless inductive power systems that could potentially be adapted for implantable devices, some of which have been shown to operate with up to 72% power efficiency [42]. The Caltech Nanofabrication lab developed one such device, a 1.4 mm by 1.4 mm glucose

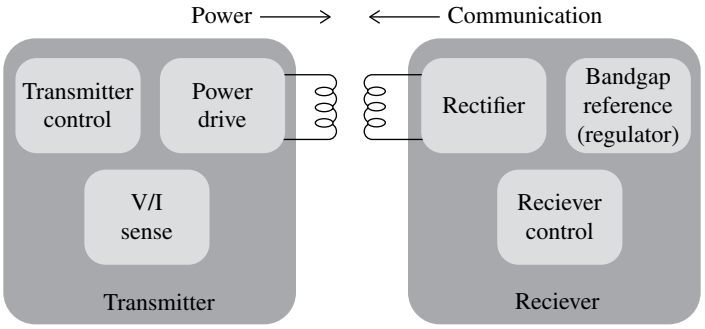


FIGURE 10.3 Basic block diagram of a four-part wireless power system.

monitoring implant that is powered by an external RF signal at 900 MHz and uses 5 μW of power [43].

Inductive power is not perfect. The coils must be matched perfectly and must be in perfect alignment with each other to realize maximum range, which is still normally only a matter of centimeters. Researchers at MIT, however, have developed a solution to these problems known as highly resonant wireless power transfer. Using this technology, they developed a system to efficiently transmit 60 W over 2 m [44]. This technology could be hugely beneficial in longer-term implants, such as cochlear implants or pacemakers, which are currently battery powered and require invasive surgery to replace the battery when it begins to fail. Highly resonant wireless power could also be used to more easily power nanomedical devices. Assuming a short lifespan for nanorobots in the body, the patient may be able to stay easily within range of a hospital transmitter implanted in their bed or wall for the entire procedure—potentially eliminating the complications introduced by the range and alignment requirements of more traditional inductive power.

Other recent examples of studies using wireless power include the design of a telemetry system based on wireless power transmission for physiological parameter monitoring [41] and a method of tracking optimal efficiency of magnetic resonance wireless power transfer system for biomedical capsule endoscopy [42]. In the former, an implanted telemetry system for experimental animals can be used to continuously monitor physiological parameters. This system is significant not only in the study of organisms but also in the evaluation of drug efficacy, artificial organs, and auxiliary devices. The system is composed of a miniature electronic capsule, a wireless power transmission module, a data-recording device, and a processing module. An electrocardiograph, a temperature sensor, and a pressure sensor are integrated in the miniature electronic capsule, in which the signals are transmitted in vitro by wireless communication after filtering, amplification, and A/D sampling. To overcome the power shortage of batteries, a wireless power transmission module based on electromagnetic induction was designed. In the latter, researchers were looking for a solution to the problem of limited battery capacity in commercialized capsule endoscopy. The chapter presents a theory for tracking the optimal efficiency of an MR-WPT (Resonant Wireless Power Transfer) system, along with its experimental

verification. A system with a 9-mm-diameter receiver is implemented, which is small enough to fit in the current capsule endoscope.

The current advances in wireless power technology are essential for the future development of nanomedical robots. However, power is only one piece of the puzzle. Until we can assemble a complete nanorobot for testing purposes, we must rely on other methods to determine how different designs interact with the human body. In Section 10.7, we will discuss one such method: computer modeling.

10.7 COMPUTER MODELING

In previous sections, we have discussed many of the basic requirements for cancer-treating nanorobots. In Section 10.7, we walk through one attempt to model all of these systems together within the human body through the use of computers.

In order to create viable nanorobots for medical purposes, all of these systems must come together in one unit. Because we are not yet able to physically produce robots on the required scale, we can use computer modeling to attempt to determine how specific combinations of all of the aforementioned systems will function once introduced into the body. In one recent example of computer modeling for nanomedicine, a research team at the University of Southern California has developed a system that models nanoscale drug delivery through the bloodstream. It has the ability to represent both active and passive targeting nanorobots. It simulates the flow of nanorobots through the bloodstream until they detect cancer cells, at which point they maneuver to the tumor and begin the process of drug delivery. This model accounts for various possible drug delivery failures, such as early release, delayed release, non-release, power failure, and immunostimulation. Upon failure, the representation of the nanorobot exits the model. Each failed delivery can be viewed in a bar graph by type of failure. When a nanorobot successfully attaches to a cancer cell, the representative nanorobot in the model incurs latency, which accounts for attaching to the cancer cell and the actual act of drug deployment. The number of successful deliveries as compared to the number of failed deliveries can be seen in another bar graph. Figure 10.4 shows an example of the included position tracking scatter plot, which displays the location of each nanoparticle or nanorobot currently in the body. The figure shows this plot immediately before nanorobots are first introduced to the body.

The model also takes into account drug toxicity and biocompatibility issues, topics discussed in Section 10.5. These topics can be impractical to test for in the lab, due to the difficulty of introducing different planes of different materials to a realistic human immune system without endangering anyone in the process. While model outputs are not as definitive as lab results, computer models can—at low costs—simulate the introduction of different materials into the human bloodstream to within an acceptable degree of accuracy. In this specific model, the body has an assumed toxicity capacity, which is the density of toxins it can contain and still be healthy. Each drug has its own toxicity level, which can be thought of as generally anti-proportional to its biocompatibility. The nanorobots release their toxins into the bloodstream every

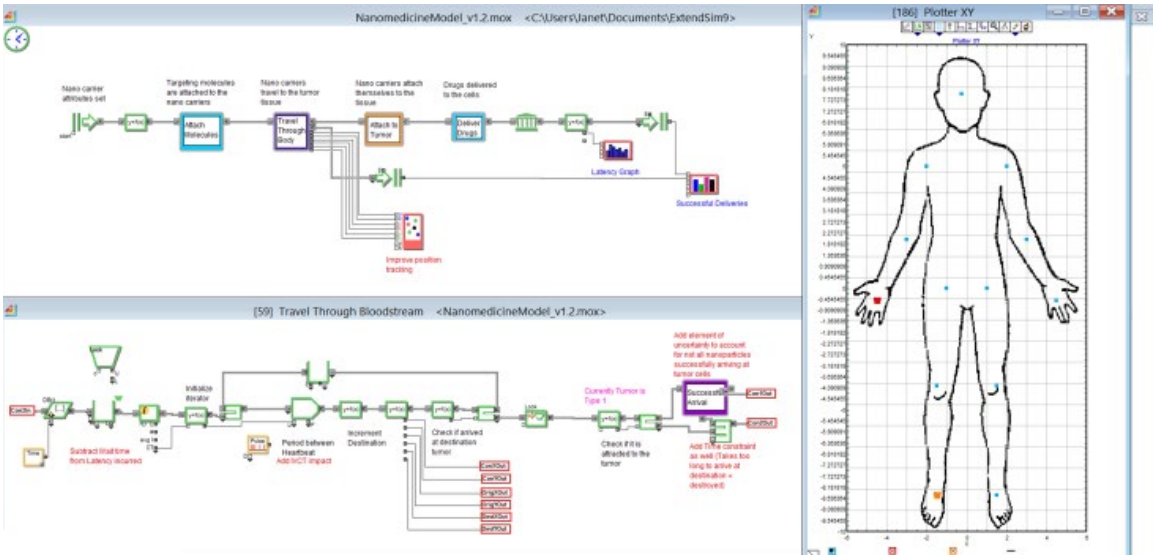


FIGURE 10.4 Position tracking scatter plot before introduction of nanorobots.

time they fail drug delivery in specific ways. This gradually increases the current concentration of toxins in the body based on the amount released, the time between subsequent releases, the distance between subsequent releases, and the toxicity value assigned to the medicine in question. Once the toxicity level is too great, the simulation ends in failure.

This model was created in ExtendSim 9, a program designed specifically for creating dynamic system models. The software was run under an educational grant provided by Imagine That!, the creators of ExtendSim. Figure 10.5 shows a flowchart of the model design within ExtendSim 9.

Data from this model can point scientists toward materials and molecular planes that are viable candidates for successful introduction to the human body, which can then be tested in the lab. This model can also predict both the overall amount and density of a specific medication can be safely released into the bloodstream before the concentration due to nanorobot failures is high enough to be dangerous to the patient.

This model has been purposely designed to grow with the increase of knowledge within the still relatively new field of nanomedicine. As more research is done and new results become available, these data can be added to the model. This will further refine the results, making this model—in theory—incredibly accurate over time. As we learn more about nanomedicine, data from this model can point scientists toward materials and molecular planes that are viable candidates for successful introduction to the human body, which can then be tested in the lab. The information collected can also help scientists develop appropriate dosages for treatment or design more efficient nanorobots with a higher delivery success rate.

10.8 RESEARCH INSTITUTIONS

In previous sections, we have explored some of the specifics of nanomedical research for cancer. In Section 10.8, we will discuss the organizations, centers, universities, and scientists actually doing the research as of the fall of 2014.

Funding for nanomedical research has been growing steadily. Research institutions around the world are receiving funding to study nanomedicine and nanotechnology. The National Institute of Health (NIH) has formed a national network of nanomedicine centers in the United States. This network comprises eight Nanomedicine Development Centers (NDCs). These NDCs were created to both advance the field of nanomedicine through research and "begin training the next generation of students in [the] emerging field [of nanomedicine] [49]."

The NIH focuses their efforts on understanding the inner workings of cells on the nanoscale and using that knowledge to "develop new technologies that could be applied to treating diseases, and/or leverage the new knowledge to focus work directly on translational studies to treat a disease or repair damaged tissue [50]." The program runs for 10 years, from 2005 to 2015. The NIH counts New York University, the University of California San Francisco, and the University of California Berkeley among its eight NDC locations. As of 2014, current research at New York University

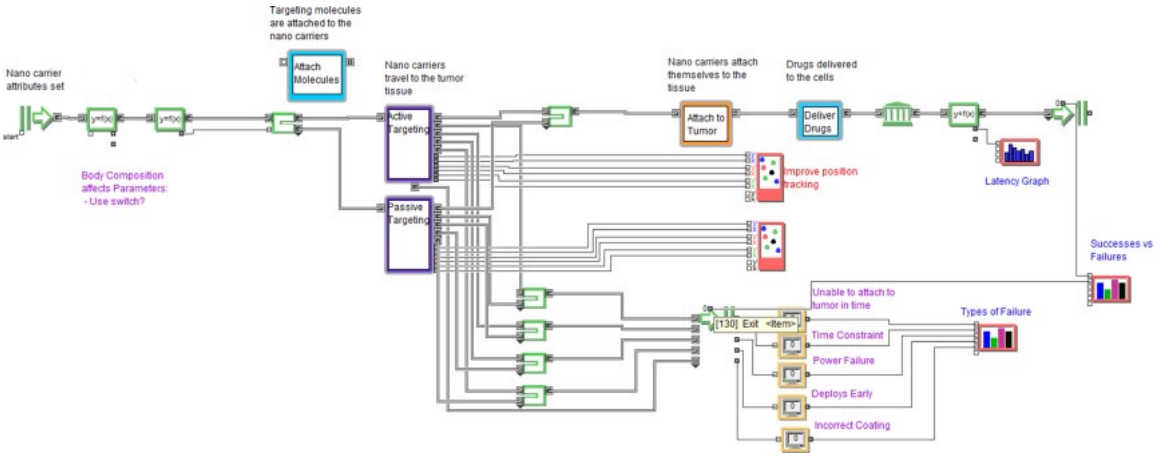


FIGURE 10.5 Flowchart of model as seen within ExtendSim 9.

includes "developing culture systems to improve adoptive immunotherapy [51]," through growing cultures of memory T cells in cultures before injecting them into the body to stimulate immune response to specific pathogens. Meanwhile, the Nanomedicine Center for Nucleoprotein Machines at Georgia Tech is researching protein creation and DNA modification with the hopes of using that knowledge to either more effectively treat or even cure sickle cell disease [52], and the University of California, San Francisco, is trying to create entirely new nanotechnological systems to diagnose and treat diseases like cancer [53].

In Europe, the European Commission created the European Technology Platform on Nanomedicine (ETPN) in 2005 with the goals of focusing research and raising money for funding. Their three priority focus areas are nanotechnology-based diagnostics including imaging, targeted drug delivery and release, and regenerative medicine [54]. Current (2014) research under the ETPN includes NANOCI, or Nanotechnology Based Cochlear Implants, which works on creating cochlear implants that connect directly to auditory nerve fibers, solving multiple problems with current assistive audio technology at once [55, 56]. Another group, NANOFOL, is working to design nanorobots to cure inflammatory diseases such as rheumatoid arthritis [55, 57].

Both the NIH and the ETPN, along with other institutions worldwide, have seen the potential for nanomedicine to forever change the landscape of modern medicine. Only with continued funding and research, however, will we develop the technology to realize the incredible potential of nanomedicine.

10.9 CONCLUSION

Although nanomedicine has the theoretical potential to repair organs, restore lost spinal function, and even reverse the aging process, for this chapter we have focused specifically on the role of nanorobots, nanoparticles, CNTs, and other nanoscale devices in the treatment of disease. Though many of the aforementioned scenarios are still merely hypothetical, modern nanomedical research provides a solid foundation for future advances in nanoscale healthcare technology.

ACKNOWLEDGMENTS

We would like to thank Samitsubhro Banerji, Yixin Cai, Saurav Chowdhury, Biyi Fang, Wanling Jiang, Krishna Suresh Reddy Padala, Padmini Prakash, Ishan Purohit, Karthik Rajamane, Yesha Shah, Yilin Xu, and Santosh Yerrapragada for their contributions to this chapter.

REFERENCES

[1] S. Sepeur. Nanotechnology: Technical Basics and Applications. Hannover: Vincentz Network GmbH & Co. KG, 2008.

[2] Center for Responsible Nanotechnology. (2008). What Is Nanotechnology?. Available at: http://www.crnano.org/whatis.htm (accessed on January 9, 2016).

[3] S. Karan and D. D. Majumder. Molecular Machinery—A Nanorobotics Control System Design for Cancer Drug Delivery. Presented at 2011 International Conference on Recent Trends in Information Systems (ReTIS), Kolkata, India, December 21–23, 2011.

[4] G. Vanderheiden. Over the Horizon: Potential Impact of Emerging Trends in Information and Communication Technology on Disability Policy and Practice. Washington, DC: National Council on Disability, 2006.

[5] M. Al-Fandi, M. A. Jaradat, M. Al-Rousan, and S. Jaradat. A Living Biological Nano Robot as Self-Navigator Sensor for Diseases. Presented at 2011 First Middle East Conference on Biomedical Engineering (MECBME), Sharjah, United Arab Emirates, February 21–24, 2011.

[6] A. A. Requicha. Nanorobots, NEMS, and nanoassembly. Proceedings of the IEEE 91(11), 1922–1933, 2003.

[7] S. P. Woods and T. G. Constandinou. Wireless capsule endoscope for targeted drug delivery: Mechanics and design considerations. IEEE Transactions on Biomedical Engineering 60(4), 945–953, 2013.

[8] Y. Zhang, D. Wang, X. Ruan, S. Jiang, and J. Lu. Control strategy for multiple capsule robots in intestine. Science China Technological Sciences 54(11), 3098–3108, 2011.

[9] S. Yim and M. Sitti. Shape-programmable soft capsule robots for semi-implantable drug delivery. IEEE Transactions on Robotics 28(5), 1198–1202, 2012.

[10] S. Sengupta, M. E. Ibele, and A. Sen. Fantastic voyage: Designing self-powered nanorobots. Angewandte Chemie International Edition 51(34), 8434–8445, 2012.

[11] K. Weidner. (2014). Nanomotors Are Controlled, for the First Time, Inside Living Cells. Available at: http://news.psu.edu/story/303296/2014/02/10/research/nanomotors-are-controlled-first-time-inside-living-cells (accessed on January 9, 2016).

[12] C. Zandonella. Cell nanotechnology: The tiny toolkit. Nature 423(6935), 10–12, 2003.

[13] S. Wang, R. Wang, P. Sellin, and Q. Zhang. DNA biosensors based on self-assembled carbon nanotubes. Biochemical and Biophysical Research Communications 325(4), 1433–1437, 2004.

[14] B. Leonard. Cancer Nanotechnology: Going Small for Big Advances: Using Nanotechnology to Advance Cancer Diagnosis, Prevention and Treatment. DIANE Publishing, Darby, PA, 2009.

[15] Y. Choi, J. Kwak, and J. W. Park. Nanotechnology for early cancer detection. Sensors 10(1), 428–455, 2010.

[16] D. Gardner and D. Shoback. Greenspan's Basic and Clinical Endocrinology. New York: McGraw Hill Professional, 2011.

[17] B. Kim, S. L. Tripp, and A. Wei. Tuning the Optical Properties of Large Gold Nanoparticle Arrays. Presented at MRS Proceedings, Boston, MA, December 2–5, 2002.

[18] A. M. Gobin, M. H. Lee, N. J. Halas, W. D. James, R. A. Drezek, and J. L. West. Near-infrared resonant nanoshells for combined optical imaging and photothermal cancer therapy. Nano Letters 7(7), 1929–1934, 2007.

[19] M. P. Melancon, W. Lu, Z. Yang, R. Zhang, Z. Cheng, A. M. Elliot, J. Stafford, T. Olson, J. Z. Zhang, and C. Li. In vitro and in vivo targeting of hollow gold nanoshells directed at epidermal growth factor receptor for photothermal ablation therapy. Molecular Cancer Therapeutics 7(6), 1730–1739, 2008.

[20] X. Huang, I. H. El-Sayed, W. Qian, and M. A. El-Sayed. Cancer cell imaging and photothermal therapy in the near-infrared region by using gold nanorods. Journal of the American Chemical Society 128(6), 2115–2120, 2006.

[21] R. Guduru, P. Liang, C. Runowicz, M. Nair, V. Atluri, and S. Khizroev. Magneto-electric nanoparticles to enable field-controlled high-specificity drug delivery to eradicate ovarian cancer cells. Scientific Reports 3, 2953–2960, 2013.

[22] P. Alivisatos. The use of nanocrystals in biological detection. Nature Biotechnology 22(1), 47–52, 2003.

[23] DARPA to Use Implantable Sensors to Monitor Soldiers in the Battlefield, 2012. Availableat:http://real-agenda.com/darpa-to-use-implantable-sensors-to-monitor-soldiers-in-the-battlefield/ (accessed on January 9, 2016).

[24] H. Lee, J. Lee, S. Oh, and C. Chung. Inductively Coupled RF Heating of Nano-Particle for Non-Invasive and Selective Cancer Cell Destruction. Presented at 10th IEEE International Conference on Nanotechnology, Islan, South Korea, August 17–20, 2010.

[25] C. J. Gannon, P. Cherukuri, B. I. Yakobson, L. Cognet, J. S. Kanzius, C. Kittrell, R. B. Weisman, M. Pasquali, H. K. Schmidt, and R. E. Smalley. Carbon nanotube-enhanced thermal destruction of cancer cells in a noninvasive radiofrequency field. Cancer 110(12), 2654–2665, 2007.

[26] P. Chakravarty, R. Marches, N. S. Zimmerman, A. D. Swafford, P. Bajaj, I. H. Musselman, P. Pantano, R. K. Draper, and E. S. Vitetta. Thermal ablation of tumor cells with antibody-functionalized single-walled carbon nanotubes. Proceedings of the National Academy of Sciences of the United States of America 105(25), 8697–8702, 2008.

[27] M. Alam and Y. Massoud. An Accurate Closed-Form Analytical Model of Single Nanoshells for Cancer Treatment. Presented at 48th Midwest Symposium on Circuits and Systems, 2005, Radisson Hotel Cincinnati Riverfront, Covington, KY, August 7–10, 2005.

[28] A. Lin, L. Hirsch, M. Lee, J. Barton, N. Halas, J. West, and R. Drezek. Nanoshell-enabled photonics-based imaging and therapy of cancer. Technology in Cancer Research & Treatment 3(1), 33–40, 2004.

[29] M. L. Brongersma. Nanoscale photonics: Nanoshells: Gifts in a gold wrapper. Nature Materials 2(5), 296–297, 2003.

[30] M. Ogiue-Ikeda, Y. Sato, and S. Ueno. Destruction of targeted cancer cells using magnetizable beads and pulsed magnetic forces. IEEE Transactions on Magnetics 40(4), 3018–3020, 2004.

[31] M. Venkatesan and B. Jolad. Nanorobots in Cancer Treatment. Presented at 2010 International Conference on Emerging Trends in Robotics and Communication Technologies (INTERACT), Sathyabama University, Chennai, India, December 3–5, 2010.

[32] R. Singhai, V. W. Patil, S. R. Jaiswal, S. D. Patil, M. B. Tayade, and A. V. Patil. E-cadherin as a diagnostic biomarker in breast cancer. North American Journal of Medical Sciences 3(5), 227–233, 2011.

[33] Biology-online.org. (2015). Life Science Reference—Biology Online. Available at: http://www.biology-online.org (accessed on June 2, 2015).

[34] Ntnu.edu. (2015). Norwegian University of Science and Technology—NTNU. Available at: http://www.ntnu.edu (accessed on June 2, 2015).

[35] J. Bonadio. Review: Local gene delivery for tissue regeneration. e-biomed: The Journal of Regenerative Medicine 1(2), 25–29, 2000.

[36] T. McKnight, A. Melechko, D. Hensley, D. Mann, G. Griffin, and M. Simpson. Tracking gene expression after DNA delivery using spatially indexed nanofiber arrays. Nano Letters 4(7), 1213–1219, 2004.

[37] G. Lalwani, A. Henslee, B. Farshid, L. Lin, F. Kasper, Y. Qin, A. Mikos, and B. Sitharaman. Two-dimensional nanostructure-reinforced biodegradable polymeric nanocomposites for bone tissue engineering. Biomacromolecules 14(3), 900–909, 2013.

[38] G. Lalwani, A. Henslee, B. Farshid, P. Parmar, L. Lin, Y. Qin, F. Kasper, A. Mikos, and B. Sitharaman. Tungsten disulfide nanotubes reinforced biodegradable polymers for bone tissue engineering. Acta Biomaterialia 9(9), 8365–8373, 2013.

[39] J. Genz, B. Haastert, G. Meyer, A. Steckelberg, H. Müller, F. Verheyen, D. Cole, W. Rathmann, B. Nowotny, M. Roden, G. Giani, A. Mielck, C. Ohmann, and A. Icks. Blood glucose testing and primary prevention of diabetes mellitus type 2—evaluation of the effect of evidence based patient information. BMC Public Health 10(1), 15, 2010.

[40] E. Jovanov, A. O'Donnell Lords, D. Raskovic, P. Cox, R. Adhami, and F. Andrasik. Stress monitoring using a distributed wireless intelligent sensor system. IEEE Engineering in Medicine and Biology Magazine 22(3), 49–55, 2003.

[41] Z. Jia, G. Yan, and B. Zhu. Design of a telemetry system based on wireless power transmission for physiological parameter monitoring. AIP Advances 5(4), 041320, 2015.

[42] K. Na, H. Jang, H. Ma, and F. Bien. Tracking optimal efficiency of magnetic resonance wireless power transfer system for biomedical capsule endoscopy. IEEE Transactions on Microwave Theory and Techniques 63(1), 295–304, 2015.

[43] D. F. Williams. On the mechanisms of biocompatibility. Biomaterials 29(20), 2941–2953, 2008.

[44] K. Riehemann, S. W. Schneider, T. A. Luger, B. Godin, M. Ferrari, and H. Fuchs. Nanomedicine—Challenge and perspectives. Angewandte Chemie International Edition 48(5), 872–897, 2009.

[45] S. Naahidi, M. Jafari, F. Edalat, K. Raymond, A. Khademhosseini, and P. Chen. Biocompatibility of engineered nanoparticles for drug delivery. Journal of Controlled Release 166(2), 182–194, 2013.

[46] N. Gómez i Bastús, E. Sánchez Tilló, S. Pujals Riatós, J. Comenge, E. Giralt Lledó, A. Celada Cotarelo, J. Lloberas Cavero, and V. Puntes. Inorganic nanoparticles and the immune system: Detection, selective activation and tolerance. Proceedings of the Society of Photo-Optical Instrumentation Engineers 8232, 823217-1–823217-14, 2012.

[47] National Cancer Institute. NCI Dictionary of Cancer Terms: Immunological Adjuvant. Available at: http://www.cancer.gov/dictionary?cdrid=43987 (accessed on January 9, 2016).

[48] National Cancer Institute. NCI Dictionary of Cancer Terms: Immunosuppression. Available at: http://www.cancer.gov/dictionary?cdrid=45727 (accessed on January 9, 2016).

[49] National Eye Institute. NIH Completes Formation of National Network of Nanomedicine Centers. National Institutes of Health, Bethesda, 2006.

[50] National Institutes of Health. (2014). Nanomedicine. Available at: https://commonfund.nih.gov/nanomedicine/index (accessed on January 9, 2016).

[51] National Institutes of Health. (2013). Nanomedicine Center for Mechanobiology Directing the Immune Response. Available at: https://commonfund.nih.gov/nanomedicine/devcenters/mechanicalbiology (accessed on January 9, 2016).

[52] Nanomedicine Center for Nucleoprotein Machines. (2010). Technical Overview. Available at: http://www.nucleoproteinmachines.org/overview.html (accessed on January 9, 2016).

[53] National Institutes of Health. (2013). Engineering Cellular Control: Synthetic Signaling and Motility Systems. Available at: https://commonfund.nih.gov/nanomedicine/devcenters/cellularcontrol (accessed on January 9, 2016).

[54] ETPN. About ETPN. Available at: http://www.etp-nanomedicine.eu/public/about (accessed on January 9, 2016).

[55] ETPN. (2014). European Funding. Available at: http://www.etp-nanomedicine.eu/public/about-nanomedicine/european-funding (accessed on January 9, 2016).

[56] NANOCI. (2013). The Concept of NANOCI. Available at: http://nanoci.org/project/index.php (accessed on January 9, 2016).

[57] NANOFOL. Scientific and Technical Objectives, January 29, 2014. Available at: http://www.nanofol.eu/index.php?id=152 (accessed on January 9, 2016).

11

NANOMEDICINE USING MAGNETO-ELECTRIC NANOPARTICLES

Mary Mehrnoosh Eshaghian-Wilner[1], Andrew Prajogi[1], Kodiak Ravicz[1], Gaurav Sarkar[1], Umang Sharma[1], Rakesh Guduru[2], and Sakhrat Khizroev[2]

[1] *EE-Systems, University of Southern California, Los Angeles, CA, USA*

[2] *College of Engineering and Computing, Florida International University, Miami, FL, USA*

11.1 INTRODUCTION

Systemic form of delivery is very effective in conveyance of the medications to the affected regions. Most of the drug administration methods depend on the circulatory system to deliver the drug to the areas of the body that need treatment. Systemic form of delivery owes its success to the vast network of blood vessels that make most cells in the body accessible. However, it is because of this very nature that the drugs reach the healthy cells as well. These drugs that are designed to treat the affected cells may or may not be benign to the healthy ones. Hence, it becomes important to find ways for selective delivery of such toxic drugs. As a solution to this problem, magnetic nanoparticles (MNs) have been annually and conventionally used for targeted delivery and release of drugs within the body, which have an edge over the systemic form of delivery. However, MNs aren't foolproof either. The application of magnetic field used to control the release of the drug attached to these MNs causes them to produce heat. Magneto-electric nanoparticles (MENs), on the other hand, do not react exothermically to the applied magnetic field and thereby are more energy efficient. To tackle this challenge, the chapter aims to discuss the MENs, their structural properties, and their application. We will also explain how MENs have better

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

energy efficiency and less heat dissipation, making them a better alternative for target drug delivery.

Controlled drug delivery technology has progressed over the past six decades. This progression began in 1952 with the introduction of the first sustained release formulation. The first generation of drug delivery (1950–1980) focused on developing oral and transdermal sustained release systems and establishing controlled drug release mechanisms. The second generation (1980–2010) was dedicated to the development of zero-order release systems, self-regulated drug delivery systems, long-term depot formulations, and nanotechnology-based delivery systems.

As the idea and applications of nanoparticles looked promising, the latter part of the second generation largely focused on studying nanoparticle formulations. The best path toward a productive third generation of drug delivery technology requires an honest, open dialog without any preconceived ideas of the past. The drug delivery field needs to take a bold approach to designing future drug delivery formulations primarily based on today's necessities and to produce the necessary innovations.

Nanoparticles can cater to the problems involved in systemic delivery and provide high drug transmission efficacy. These nanoparticles include MNs and MENs that can be controlled by magnetic fields. A major difference between MNs and MENs is that the magnetic field of the latter changes the internal electric fields which exists between the bonds. In this chapter, three experiments are discussed in detail. The first deals with the application of MENs in the treatment of ovarian cancer, the second relates to the delivery of drugs for treatment of HIV-1 reservoirs, and the third explains how magneto-electric (ME) spin can be used to stimulate the brain.

Section 11.6 talks about bioceramic nanoparticles and their integration with MNs. We will see how they are prepared and the important characteristics of such hybrid nanoparticles. We will also delve into how these so-called mesoporous particles react to external stimuli to release the encapsulated drugs in a controlled way. We will then propose to substitute the MNs used in the core of these nanostructures with MENs because of their superiority over the conventional MNs.

11.2 OVERVIEW OF MENs

MENs exhibit multiferroic [1] characteristics, which implies not only the magnetization property but also the dielectric polarization can be affected by magnetic field. They consist of magnetic nanofibers like $CoFe_2O_4$ combined with ferroelectric nanofibers like $BaTiO_3$ [1]. At nano level, quantum mechanics comes into action, which results in coupling between the magnetic spin and the electric dipole.

11.2.1 How Do They Function?

The ME effect helps in controlling the ionic bond between the drug and the MEN particles. The ME effect can be controlled by an externally applied alternating current (a.c.) or direct currect (d.c.) magnetic field. This same ME effect of the MENs can be

used to create electric fields near cell membranes, thereby altering the size of the pores of the tumorous cells and increasing the intake of drug by the cells. So to summarize how MENs serve as perfect candidates for drug delivery, one can describe it's functionality in the following manner:

1. *Transmission and targeting*: The drug-loaded MENs can be directed to a specific region or cells within the body using a d.c. magnetic field. They flow with the body fluids, but their motion can be controlled by applying d.c. field of the right intensity. Thus the goal of selective targeting can be achieved using MENs.
2. *Drug release*: The need for on-demand release of the drug by the MEN is made possible due to the ME property of the MEN. By applying an alternating magnetic field of the correct magnitude, release efficacy can be achieved. This concept is explained more elaborately in Section 11.3.
3. *Drug intake*: The MENs are also a valuable asset when it comes to increasing the drug intake capacity of cancerous cells. It makes use of the difference in electrical properties of cell membranes (difference of properties between cancerous and noncancerous cells) to release drugs only into the cancerous cells leaving the healthy ones unharmed. More on this application is described in Section 11.4.

Thus, we can see why the MENs have great potential as a highly selective drug delivery system.

11.2.2 MNs vs MENs

MNs versus MENs—While the ME effect in case of MENs take place in an energy-efficient manner without resulting in any heat dissipation, in case of MNs it is not the same. The drugs adhered to the MNs get released either due to the heat developed on the application of the external magnetic field or mechanical forces produced due to the applied field. In both the cases, the cause of separation is extrinsic to the nanoparticles. However that is not the case with MENs. In MENs, no heat or additional particles are produced. Therefore, they have better energy efficiency and less heat dissipation. Now that the differences are clear, we will be focusing on how these MENs were produced.

11.3 EXPERIMENT 1: EXTERNALLY CONTROLLED ON-DEMAND RELEASE OF ANTI-HIV DRUG AZTTP USING MENS AS CARRIERS

Ever since the first reported case of HIV, a substantial amount of effort has been put into the research and development of antiretroviral (ARV) drugs. ARV drugs, also known as anti-HIV or anti-AIDS drugs, aim at limiting the amount of HIV in the body to a lower level. This strengthens the immune system of the patient and allows it to recover from any damage that may have been caused by the HIV virus. ARV

drug therapy is the main type of treatment for the HIV virus and AIDS. Although it is not a permanent cure, it prolongs the life of the patient by keeping illness at bay for many years. The treatment requires the patients to take drugs every day for the rest of their lives.

Although highly active ARV therapy has resulted in remarkable decline in the morbidity and mortality in AIDS patients, inadequately low delivery of ARV drugs across the blood–brain barrier (BBB) results in virus persistence. The capability of high-efficacy targeted drug delivery and on-demand release remains a formidable task. Here, we report an *in vitro* study to demonstrate the on-demand release of azidothymidine 5′-triphosphate, an anti-HIV drug, from 30-nm $CoFe_2O_4$–$BaTiO_3$ MENs by applying a low a.c. magnetic field. MENs as field-controlled drug carriers offer a unique capability of field-triggered release after crossing the BBB. Due to the intrinsic magnetoelectricity, these nanoparticles can couple external magnetic fields with the electric forces in drug–carrier bonds to enable remotely controlled delivery without exploiting heat. Functional and structural integrity of the drug after the release was confirmed in *in vitro* experiments with HIV-infected cells and through atomic force microscopy (AFM), spectrophotometry, Fourier transform infrared (FTIR) spectrometry, and mass spectrometry studies.

Often a "combinational therapy" is touted over the one-drug therapy since in the latter case, the HIV becomes resistant to the drug, and hence its effectiveness over time reduces. On the other hand, taking two or more ARVs at a time vastly reduces the rate at which this resistance to the drugs would develop, thereby ensuring the efficacy of the drugs lasts over a longer period of time. This combinational therapy involving the use of multiple anti-HIV drugs at the same time is sometimes referred to as highly active antiretroviral therapy (HAART). Although HAART has resulted in remarkable decline in the morbidity and mortality in AIDS patients, zero or inadequately low delivery of ARV drugs across the BBB and into the brain and other tissue organs results in virus persistence [2–4]. Hence, the elimination of HIV-1 reservoirs still remains a formidable task [4, 5]. In recent years, the use of nanotechnology in medicine has shown exciting prospect for the development of novel remotely controlled drug delivery systems [6, 7]. Besides MNs [8–10], other delivery systems rely on using thermally responsive polymers, optically (ultraviolet, visible wavelength, and IR) and acoustically activated nanostructures, liposomes, electrochemical processes, and others. An extensive review of these and other active triggering approaches was presented in the article by Timko et al. [11]. The features of magnetically forced triggering give this approach unique advantages over other similar techniques.. In this case, the speed of delivery is determined by the external magnetic field. For instance, as it is related to the delivery across the BBB, this approach provides a way to deliver drugs sufficiently fast to avoid their engulfing by the reticuloendothelial system (RES). Nucleotide reverse-transcriptase inhibitor (NRTI) 3′-azido-3′-deoxythymidine-5′-triphosphate (AZTTP) is among the most challenging ARV drugs to deliver across the BBB [10, 11].

11.3.1 Hypothesis

Previous researches dealt with the applications of MNs tagged with AZTTP. Application of external magnetic fields resulted in transmigration of these MNs across the BBB without affecting the integrity of the BBB. The results demonstrated that these MNs acted as inhibitors to HIV-1 p24 antigen production in an *in vitro* infection model system compared to the free AZTTP. The down point of this method proved to be the uncertainty of drug release from the carrier if and when the nano-carrier reaches the target. The current presumptive mechanisms of drug delivery depend on manually uncontrollable cellular phenomena such as exocytosis of drug containing intracellular vesicle, intracellular Ca^{2+} concentrations and pathology-specific responses (change in pH, temperature, etc.) [12]. As a result, the binding force between the nano-carrier and the drug must be maintained relatively weak in order to obtain a relatively small physiologic change and ensure the release of the drug. Consequently, more than 99% of the drug/carriers are deposited in the liver, lungs, and other lymphoid organs before they reach the target. Therefore, an approach for finely controlled and enhanced release of AZTTP in sufficient therapeutic levels in the brain or in other target organs is still sought after, as it is critical for the complete eradication of HIV.

We present a study to demonstrate that dissipation-free, energy-efficient and low-field on-demand drug release can be achieved if the conventional MN carriers are replaced by MENs [13]. ME materials represent a relatively recently introduced class of multifunctional nanostructures in which magnetic and electric fields can be strongly coupled at body temperature [14]. Similar to the MNs, MENs can be designed to have adequately high magnetic moments and, therefore, also can be used for targeted delivery by applying remote d.c. magnetic fields. However, due to their nonzero magneto-electricity, unlike the traditional MNs, MENs offer an additional feature that can enable a new dissipation-free mechanism to force a high-efficacy externally controlled drug release process at the sub-cellular level using remote low-energy d.c. and/or a.c. magnetic fields.

An exaggerated illustration in figure 1 [12] explains the concept of the field-triggered on-demand drug release from MENs. Let's simplify the description by using an example with a remote magnetic field in one specific direction, that is, along the x-axis, with respect to the MEN-drug nano-complex. The original (zero-field) ionic bond, with charge Qionic of the nanoparticle, is schematically shown in figure 1a [12]. The AZTTP molecules (typically interconnected in chains) surround each MEN in a symmetric fashion. As shown in figure 1b [12], the application of a nonzero magnetic field results in the formation of a nonzero electric dipole moment in the nanoparticle due to the nonzero ME effect. For simplicity assuming an isotropic model, the triggered dipole moment $\Delta P = \alpha H$, where α is the first-order ME coefficient and H is the magnetic field. The amplitude of the dipole charge surface density on each side of the nanoparticle along the field would be of the order of $\sigma_{ME} \sim \pm \alpha H$, where "positive" and "negative" signs are applied to the opposite sides of the dipole, respectively. The dipole moment breaks the original symmetry of the charge in the

MEN shell. Consequently, as the magnitude of the magnetic field is further increased above the threshold value at which the dipole charge density on the "negative" side becomes comparable to the positive ionic charge density in the shell, $\sigma_{ME} \sim Q_{ionic}/\pi d^2$, that is $H_{th} \sim Q_{ionic}/\pi d^2 \alpha$, where d is the diameter of the MEN, the bond in this direction along the x-axis will be broken, while the opposite bond will be further strengthened, as illustrated in figure 1c [12]. By symmetry, to break the bond in the opposite direction, the field sequence should be repeated in the reverse direction, as illustrated in figure 1d and e [12]. This simplified scenario doesn't take into account the randomness of the orientations of the nanoformulations. Ideally, applying an a.c. magnetic field that equivalently sweeps all bond orientations will create a more uniform and efficient bond-breaking process over the surface of the nanoformulation and thus enhance the drug release efficacy. In the next generation of the technology, this goal can be achieved by using a spatially rotating field, which in turn can be accomplished, for example, by using an array of coils that generate a.c. fields with nonzero phase shifts with respect to each other.

11.3.2 Experiment

This hypothesis was verified in this study by demonstrating the on-demand release of AZTTP from 30-nm $CoFe_2O_4$–$BaTiO_3$ MENs by applying low a.c. and d.c. magnetic fields. It was seen that a 44-Oe a.c. field at 1000 Hz was sufficient to trigger over 89% release. Further, a 66-Oe d.c. field to trigger a comparable release level could be used to support the dissipation-free release model.

11.3.3 Results

We used AFM in conjunction with UV-spectrophotometry, FTIR analysis, and mass spectroscopy to directly trace the kinetics of the drug release process at different stages of the release under the influence of remote d.c. and a.c. magnetic fields. The three key stages included:

I. the **initial state** with separate MEN carriers and AZTTP molecules
II. the **loaded state** in which MEN-drug nano-formulations are formed
III. the **final state** after the a.c.-field-forced separation of AZTTP and MENs, that is, after the on-demand drug release.

In vitro experiments on HIV-infected human cells were conducted to demonstrate the structural and functional integrity of AZTTP after this physical release process. In the following, we present the key results of this experimental study. The experiments are described in more detail.

11.3.3.1 Transmission Electron Microscopy Study of MENs Transmission electron microscopy (TEM) is a technique where electrons, instead of light, are used for viewing things at the nano level. This is possible because of the small De Broglie wavelength of electron as compared to light. In the experiments described later, for the role

of MENs we used nanoparticles made of the popular core-shell composition $CoFe_2O_4$–$BaTiO_3$, in which the relatively high moment $CoFe_2O_4$ 1-nm shell was used to enhance the ME coefficient [15]. In general, nanoparticles as small as 5-nm in diameter can be fabricated with physical methods such as ion beam proximity lithography or imprint lithography [16]. In this study, considering the novelty of the approach, we focused on the main discovery of using MENs for on-demand drug release rather than on the development of scaling approaches. The default measurements were conducted with 30-nm MENs. A typical TEM image of the fabricated MENs, with clearly visible core-shell structures, is shown in figure 2(a) [12]. The composition of the MENs was confirmed through energy-dispersive spectroscopy (EDS), as shown in figure 2(b) [12]. The ME coefficient for the nanoparticles was measured via point I–V methods in the presence of a field to be the order of 100 V/(cm Oe).

11.3.3.2 Spectrophotometry Study of the Release The goal of the first experiment was to measure the amount of the drug (AZTTP) at different stages of the release process using the conventional approach of spectrophotometry by measuring the UV-light absorption at the 267-nm maximum of the drug's absorption spectrum. We could bind approximately 24% of the drug to the nanoparticles by incubating AZTTP with the MENs in the Tris-EDTA (TE) buffer (pH 7.4) for 3 h. The amount of the bound drug was determined by estimating the concentration of AZTTP in the unbound fraction (supernatant) of the incubation mixture by spectrophotometry. A low-energy low-field Helmholtz pair connected to a function generator was used to apply remote d.c. and a.c. magnetic fields. The chart in figure 3 [12] summarizes the key results of this experiment, in which we measured the amount of the unbound drug depending on the external field strength and frequency. It can be seen that application of a 44-Oe field at 1000 Hz results in 89.3% of the drug being released compared to only 16.4% of the drug released upon application of a d.c. field of the same amplitude. The data also confirms that a lower magnetic field with higher frequency is needed to break the bonds, because oscillation of the bonds caused by a higher frequency field facilitates the breaking of the bonds. However, when a strong enough d.c. field (> ~65 Oe) is applied, the bond-breaking side of the drug chain can gain enough momentum, causing the chain to break free from the MEN shell even at zero frequency. To fully exploit the potential of the new controlled drug release nanotechnology, it is important to conduct a more detailed pharmacokinetics study.

11.3.3.3 Pharmacokinetics Study In this experiment, MENs loaded with AZTTP drug were exposed to an external magnetic field at different strengths (12, 44, and 66 Oe) and frequencies (0, 100, and 1000 Hz) at different treatment durations (1, 5, 10, 60, and 120 min) in order to understand the release kinetics. Results are summarized in the three dimensional (3D) chart in figure 4 [12]. For every field–frequency combination set, we used a fresh solution with AZTTP-loaded MENs.

11.3.3.4 FTIR Analysis The concept of the drug release by a remote magnetic field was confirmed also through FTIR analysis, as shown for the three key stages of the process kinetics in figure 5 [12]: (top) free MENs and AZTTP before loading,

(middle) loaded state: MEN-AZTTP nanoformulations, and (bottom) MENs after the AZTTP release by a remote 44-Oe field at a 100-Hz frequency. Compared to the initial and final unbound states, the loaded state showed almost 30% weaker absorbance in general and a transformed spectrum in the wavenumber region from 1750 to 1250 cm^{-1}.

To observe the release process at the molecular level, we conducted the following AFM measurements. Figure 6 [12] shows a sequence of AFM images that reflect the following four stages of the release process: free (a) MEN and (b) AZTTP chains, (c) loaded MEN-AZTTP nanoformulations, and (d) MENs and (e) AZTTP after the release by a 44-Oe a.c. field at 1000 Hz. To obtain images c and d, the unloaded drug was washed away with the supernatant. One can observe that the MENs and AZTTP chains before the loading step and after the release process look similar.

11.3.3.5 Field-Controlled Delivery and Drug Release by MENs Loaded with AZTTP across BBB in* In Vitro *Model Translocation experiments were performed on day 5 of BBB cell culture. In order to achieve the translocation of AZTTP across the BBB, AZTTP were loaded on MENs and subjected to an external magnetic field of 40 Oe (to avoid any unwanted drug release at higher fields) for 3 h and a gradient of approximately 22 Oe/cm (to pull the nanoparticles across the BBB) for 6 h of the incubation period. To apply a field normal to the BBB, the field coils were placed below the cell culture wells carrying the BBB model. Once the incubation was completed, the medium in the bottom chamber was isolated and subjected to the a.c. magnetic field (66 Oe at 100 Hz field for 5 min) to increase the release percentage. After the field treatment, the MENs were isolated via a magnetic separation technique. Once separated, the supernatant was measured for the drug concentration via spectrophotometry. The results indicated that approximately 40% of the drug was translocated across the BBB as a result of the process.

11.3.4 Discussion

The spectrophotometry experiment (fig. 3) indicates the threshold value that is necessary for the field to start the release of the AZTTP drugs by the MENs, to be of the order of 10 Oe, even though only 10% of the drug was released at such a low field value. Using the simplified expression we derived before for the release threshold field ($H_{th} \sim Q_{ionic}/\pi d^2 \alpha$) and assuming the nanoparticle diameter $d = 30$ nm and the ME coefficient $\alpha = 100$ V/(cm Oe), we can evaluate the value for the effective charge, Q_{ionic} (responsible for the ionic/covalent bond between the MEN and AZTTP cluster), to be approximately 10^{-17} C. The estimate seems reasonable and indicates that approximately a few hundred electrons are involved in the bond formation. Further, in agreement with the original hypothesis, the experiment also proves that the a.c. field provides a substantially more enhanced release compared to that of the d.c. field. Indeed, at 12-Oe field amplitude, d.c. and a.c. (100 Hz) fields results in 1.9 and 10% released drug percentage, respectively. To achieve almost full (89.3%) drug release, a field of 44 Oe at a frequency of 1000 Hz is required. The effect of the a.c.-field control is significant in this case. For comparison, with a field of 44 Oe at the frequency of

100 Hz, the released amount was 28.5% while for the d.c. case at the same field amplitude was only 16.4%. The experiment indicates that such nanoformulations can be directed to the targeted cells via a d.c. magnetic field with a spatial gradient while the drug can be effectively released on demand via an external a.c. magnetic field at 1000 Hz. Note that the highest applied field of 65 Oe immediately saturated the system resulting in an almost full drug release even in the d.c. case. According to our hypothesis, the latter can be explained by the magnetic-field-triggered large electric force that gave enough momentum to break the bond. The results of the more detailed pharmacokinetics study (fig. 4) also indicate the presence of a certain threshold field that needs to be applied to trigger the release. The threshold values at different frequencies reach their saturation value in about 60 min of the incubation process. Again, this threshold field exists even in the d.c. case (zero frequency) and can be further reduced in the a.c. case by increasing the frequency of an a.c. field. To separate the d.c. and a.c. functions for drug delivery and release, respectively, it is important to maintain the amplitude of the field below the saturation value, for example, at 44 Oe. The FTIR results are in agreement with spectrophotometry data (fig. 5). Supporting the earlier release model, the initial state (before drug loading) and the final state (after the release) FTIR spectra look similar while the altered intermediate state reflects the effect of binding between MENs and AZTTP molecules. The AFM measurements (fig. 6) directly illustrate the physical difference between MEN-AZTTP nanoformulations and the free MENs before loading and after releasing the drugs. In agreement with the described spectrophotometry and FTIR measurements, the AFM analysis also indicates that the MENs return to their original unbound state after the drug release triggered by a.c. magnetic field. In addition, AFM also proves the nanoscale nature of the approach, with the potential to be applied at the subcellular level. The *in vitro* experiment using human PBMCs infected with HIV-$1_{Ba\text{-}L}$ proves the intact integrity of the drug after the a.c.-triggered release [12].

11.4 EXPERIMENT 2: MENS TO ENABLE FIELD-CONTROLLED HIGH-SPECIFICITY DRUG DELIVERY TO ERADICATE OVARIAN CANCER CELLS

By exploiting the difference in the electric properties of the membrane between the tumor and healthy cells and utilizing the characteristic property of MENs to serve as nanosized converters of remote magnetic field energy into MEN's intrinsic electric field energy, we were able to achieve high-specificity targeted delivery of anti-neoplastic drugs in the treatment of cancer (particularly ovarian cancer). Once the localized nano-electroporation sites have been created, we can trigger high-specificity drug uptake by remotely controlling the membrane electric fields.

These tests were carried out *in vitro* on human ovarian carcinoma cell (SKOV-3) and healthy cell (HOMEC) lines. A 30-Oe d.c. field was applied to trigger high-specificity uptake of paclitaxel loaded on 30-nm $CoFe_2O_4$–$BaTiO_3$ MENs. The results showed that the drug penetrated through the membrane and completely eradicated the tumor within 24 h without affecting the normal cells.

The past few decades have seen a dramatic change in the methodology of drug delivery, adapting a more engineered technology for the targeted delivery and controlled release of therapeutic agents. Transportation through circulatory system or through cells and tissues, although effectively delivers the drug to the carcinogenic cells, its effect on the surrounding healthy cells is often responsible for producing severe side effects and toxicity. These side effects, such as severe nausea, upset stomach, and unwanted drowsiness, have proven to be a hindrance toward producing optimal medication for treatment of many carcinogenic and neurodegenerative diseases. Delivering the drugs locally rather than systematically through the whole body is now being used as a common way of reducing these side effects and achieving an increased impact of the drug on the target cells. One such targeted delivery technique that has shown improved survival rates is intraperitoneal (IP) delivery in ovarian cancer. In IP therapy, anticancer medication is mixed with fluids and injected through an access port and catheter directly into the peritoneal space. The medication comes into direct contact with the cancer, allowing a higher concentration to be given to the patient. However, catheter complications and toxicity have precluded widespread adoption of this invasive means of delivery. Current research attempts try to deal with these issues by adopting nanoscale systems. Monoclonal antibodies are used to recognize the tumor-specific biomarker, while the nanoscale control further improves the specificity and targeted drug delivery capability in general.

One of the most important challenges in research is the crossing of BBB. It is a barrier with highly selective permeability that regulates the transfer of essential substance through the bloodstream into the central nervous system (CNS). It also prevents the entry of substances that may harm the brain. The following study addresses this challenge through a new physical concept based on the electrical properties of the membranes and the characteristics of MENs.

11.4.1 Hypothesis

Similar to MNs, MENs have a nonzero magnetic moment and, therefore, can be controlled remotely via application of an external magnetic field. The difference arises from the fact that an external magnetic field is used in MENs, giving them a greater functionality through the use of energy-efficient controls of the nanoparticle intrinsic electrical fields. This unprecedented capability is a result of the strong ME coupling in this new class of nanostructures even at body temperature. As a result, MENs introduced in a biological microenvironment act as localized magnetic-to-electric-field nano-converters that allow remote control and generation of the electric signals that underlie the intrinsic molecular interactions. This capability was exploited to achieve remotely controlled brain stimulation in patients with Parkinson's disease by applying low-energy a.c. magnetic fields to control the a.c. electric signals in the central nervous system (CNS) using intravenously injected MENs and to deliver and release on-demand (via an external field) ARV drug AZTTP for treatment of HIV-1 reservoirs across the BBB. In this section, we discuss the field-controlled specificity of the drug-loaded MENs as required to significantly improve the state of chemotherapy.

The cell membrane can be thought of as an electrically polarizable medium because of the presence of ion channels and other electric-field-driven properties. As a result, its properties can be significantly affected by an electric field. One such characteristic that exploits this dependence of the membrane's porosity on the electric field is electroporation.

The electroporation has been widely studied as a means to trigger drug delivery into the cells. It was learned through macroscale studies that an electric field of higher than 1000 V/cm creates sufficiently large pores for the drug nanoformulations to penetrate through the membrane. The approach in this study was to scale down this process to the nano level by using MENs to exploit this promising drug delivery technique. This process, called nano-electroporation, involves magnetic-field-activated MENs loaded with the drug and optionally with the biomarker-specific antibodies (for delivery to the tumor cells) can generate localized fields large enough to open up the membrane pores in their proximity only and thus let the drug inside the tumor cells. Most of the energy is consumed in the opening up of the local pores, which forms the main operation of the process. Thus the process proves to be energy efficient in the sense that it does not result in any significant and potentially damaging energy dissipation, for example, in terms of heat. The interaction between the MENs and the electric system of the membrane can be effectively summed as behaving like a field-controlled gate to let the drug-loaded nanoparticles enter specifically the tumor cells only.

Figure 7 [17] shows the main hypothesis. It can be seen that the origin of the specificity to the tumor cells can be analyzed as a two-step process. In the first step, the biomarker-specific antibodies navigate the drug-loaded MENs (to which they also are attached) to the tumor cell membrane. The second step is driven by the difference in the threshold field values, H_{th}, for the "gate" to open up, thereby assisting in achieving even higher specificity. Indeed, it is well known that the electric properties differ significantly between the healthy and tumor cells of the same type 22. It has been observed that, in general, the tumor cells have substantially lower values of the potential compared to that of the healthy cells.

The cancer cells thus have a significantly lower value of the threshold field for the drug-loaded MENs to enter the cell. Considering the value for the ME coefficient $\alpha \sim 100$ mV/(cm Oe), according to the simple isotropic expression for the ME effect, $\Delta P = \alpha H$, where P and H stand for the induced electric dipole field and the external magnetic field, respectively, the electric field of the order of 1000 V/cm can be generated a few nanometers away from the MEN merely by applying a magnetic field of 10 kOe. Moreover, it is possible to generate electric fields of the same order in magnitude by much smaller magnetic fields, of the order of 100 Oe, if one takes into account the pyramidal shape of the real-life nanoparticles, as shown later, because of the high-density charge accumulation at the edges. Once the drug-loaded MENs have penetrated into the cell cytosol through the "open" pores in the membrane, the magnetic field is further increased beyond a second critical value, H_r, necessary for overcoming the drug-MEN binding energy, ultimately causing the drug to be released off the MENs.

One underlying hypothesis of this field-controlled drug release process that came into consideration recently is that this field strongly depends on the binding force

between the MEN and the drug. As a consequence, different intermediate coating materials, field excitation frequencies, and treatment durations can be used to tune the field in a large range. To summarize, in the ideal case there are two critical field values:

1. H_{th}: defines the drug penetration threshold through the tumor cell membrane and
2. H_r: the following release of the drug into the cell cytosol.

The condition to ensure adequately high efficacy of the uptake is $H_r > H_{th}$. While the condition to be met in order to ensure the required specificity of the uptake to the cancer cells only, the external applied field, H_A, needs to be higher than the release field for the tumor cells, Hr_cancer, and lower than the threshold field for the healthy cells, Hth_healthy.

Thus MENs can not only provide field-controlled delivery but also can significantly improve the specificity to tumor (compared to the specificity defined by the monoclonal antibodies alone). In fact, the combination of MENs and monoclonal antibodies can further result in an even better delivery system in which the monoclonal antibodies drive the loaded drugs toward the surface of the tumor cells and the field-controlled MENs move the drugs across the cell membrane into the cytosol.

11.4.2 Experiment

This study puts the aforementioned hypothesis to test in the treatment of cancer, specifically using epithelial ovarian cancer (EOC), which has been widely studied in the medical community. Cytoreductive surgery followed by chemotherapy with mitotic inhibitor Paclitaxel (PTX) with platinum is the gold standard in treating EOC. The drug administration is intravenous (IV) in the majority of the cases and IP in some cases. Although IP therapy is more effective than IV, its application is still limited owing to the side effects discussed earlier in this chapter. Both the therapies offer lesser specificity of the drug uptake, and as a result EOC remains a highly lethal malignancy, which underlines the relevance of the current study to this field. Because of the high-specificity capability, the new nanotechnology can be used for targeted treatment of both localized and metastasized tumor cells. Finally, this nanotechnology can be further applied to a wide range of other cancers.

11.4.3 Results

Through the described *in vitro* studies on human ovarian carcinoma cell (SKOV-3) and healthy ovarian cell (HOMEC) lines, we demonstrated that high-specificity uptake of PTX-loaded 30-nm $CoFe_2O_4$–$BaTiO_3$ MENs could be triggered by a low-energy 30-Oe d.c. remote magnetic field with negligible heat dissipation [18–20]. Through kinetics studies, we confirmed that the drug penetrated through the tumor cell membrane and eradicated the majority of the cells within a 24-h period without affecting the surrounding healthy cells. Finally, to demonstrate the applicability of

this nanotechnology to other cancers, we conducted a parallel study using a multi-drug-resistant (MDR) uterine sarcoma cell type MES-SA/DX5 [21–23].

The release threshold field, H_r, could be controlled in a wide range, from 10 Oe to substantially over 200 Oe, through different intermediate layers/coatings. By default, in order to provide adequate coupling between the MENs and Flutax-2 (to provide the initial release field of the order of 30 Oe), before being loaded with the drug, the MENs were coated with 3-Å thick glycerol monooleate (GMO) layers. For the purpose of a comparative analysis, we studied the following combinations of nanoparticles: (i) MENs loaded with PTX, (ii) MENs loaded with PTX and the popular cancer biomarker HER-2 antibody, (iii) free PTX, and (iv) conventional MNs loaded with PTX. As the conventional MNs, 30-nm magnetite nanoparticles were used [17].

11.4.3.1 Field-Controlled Drug Release by MEN-Based Carriers Drug release from these different MEN-based combinations was triggered by a magnetic field at different strengths and frequencies. The pellet obtained after the drug loading procedure was washed thrice with the phosphate-buffered saline (PBS) buffer, to remove any residual unbounded drug. The drug-loaded-MENs' pellet was added to 1 ml of the PBS buffer in a vial and subjected to a d.c. or a.c. magnetic field using a pair of Helmholtz coils connected to a d.c. or a.c. power supply, respectively. After exposing the vial to any magnetic field environment under study, the supernatant was obtained by spinning the sample at 3000 rpm for 5 min and at 10°C. The supernatant was measured for the amount of the released drug spectrophotometrically through the absorbance at the PTX maximum wavelength of 230 nm [24].

The results of the field-controlled drug release spectrophotometry (absorption) experiments are summarized in figure 8 [17]. Figure 8A [17] shows the percentage of the drug release after a 1-min exposure to a magnetic field at three strengths, 12, 44, and 66 Oe, respectively, for three different frequencies, 0, 100, and 1000 Hz, respectively. As expected (see earlier explanation), for each frequency, there was a critical field, H_r, at which the drug release was significantly boosted. The increase of the frequency in the range up to 1000 Hz under study increased the release efficacy (by over 40%), especially at the low field range. Figure 8B [17] illustrates the kinetics of the field strength frequency dependence of the release for the five values of the field exposure times, 1, 5, 10, 60, and 120 min, respectively. The quantitative values are also presented in Table S4. For every exposure time setting, a fresh solution with PTX-loaded GMO-coated MENs was used. The field-triggered drug release was also confirmed through AFM, FTIR, mass spectrometry, and X-ray diffraction (XRD) pattern studies [24].

11.4.3.2 Field-Controlled Drug Uptake by Tumor Cells through the MEN-Induced NANO-ELECTROPORATION Fluorescent cellular drug uptake experiments were performed on the SKOV-3 cells using the four different drug forms under study, (i) free Flutax-2, (ii) Flutax-2 bound to the conventional MNs, (iii) Flutax-2 bound to GMO-MENs, and (iv) Flutax-2 bound to HER-2-GMO-MENs, respectively. The obtained Flutax-2 concentration was normalized to the protein amount. The results of

the experiment performed in triplicates are shown in figure 9 [17]. These results showed that the drug uptake increased by a factor of 5 for the drug carried by field-controlled MENs compared to the drug driven by the HER-2 antibodies.

11.4.3.3 Confocal Microscopy to Visualize the Internal Drug Localization in SKOV-3 Cell Lines To visualize the internal localization of each of the four drug forms under study, (i) free Flutax-2, (ii) Flutax-2-GMO-MENs, (iii) Flutax-2-HER-2, and (iv) Flutax-2-MNs, respectively, in SKOV-3 cell lines, we conducted the following fluorescence imaging experiments.

11.4.3.4 Magnetic Field Dependence of Drug Uptake in Cancer and Healthy Cells To understand the field dependence of the described process, we performed the cellular drug uptake experiments under a varying magnetic field strength on both cancer ovarian (SKOV-3) and healthy ovarian (HOMEC) cell lines. The HOMEC cells were cultured according to the same procedures that are described for the SKOV-3 cells in Section 11.5.1. As a control, the cells with GMO-MENs only (without Flutax-2) were treated under the equivalent conditions. The cell culture plates with the MENs and drug-GMO-MENs were exposed to three different field strengths, 5, 15, and 30 Oe, respectively. The results are summarized in figure 11 [17]. The measurements showed that as the field was increased above approximately 30 Oe, the drug penetration into the cancer cells (SKOV-3) greatly increased. On the other hand, it can be noted that the drug barely affected the healthy cells (HOMEC) in the field range under study.

11.4.3.5 Cancer Cell Viability Test After we confirmed that the drug-loaded MENs in the vicinity of the cancer cells indeed acted as a field-controlled valve to let the drug in (due to the effective nano-electroporation effect according to our hypothesis), we studied the viability of the cancer cells for different combinations of the drug and the carrier after the drug penetrated through the cell membrane. (Here, maintaining the remote field at 30 Oe provided the specificity to the cancer cells or, in other words, ensured that the healthy cells were intact.) The confocal images that were obtained after a 24-h field treatment are in figure 6 [12]. The three key combinations of the carrier included (i) no particle; (ii) HER-2-GMO-MENs (note: here, HER-2 stands for the HER-2 biomarker antibody); and (iii) GMO-MENs, respectively. Accordingly, the three images (from left to right) in figure 12A [17] show the morphology of the cancer cells after 24-h treatment by (i) the free drug (with no particle carrier), (ii) drug-HER-2-GMO-MENs with no field applied, and (iii) drug-GMO-MENs in a 30-Oe d.c. field. The three control images in figure 12B [17] show the morphology of the cancer cells after the 24-h treatment by the same three combinations of the carrier with no drug present. In addition, we conducted the confocal imaging and the trypan-blue cell viability tests on both SKOV-3 and HOMEC cell lines after 24- and 36-h field treatment.

11.4.3.6* In Vitro *Cytotoxicity Assay To determine the cytotoxicity of the GMO-MENs on SKOV-3 cells, a quantitative colorimetric XTT (sodium 2,3-bis-(2-methoxy-4-nitro-5-sulfophenyl)-5-[(phenylamino)-carbonyl]-2*H*-tetrazolium

inner salt) assay was performed. The assay is based on the reduction of XTT tetrazolium salt by the viable cells to form orange-colored formazan derivative. In this assay, 1×10^5 cells were seeded per well in a 96-well plate and incubated at 37°C for 24 h. After the incubation, the cell medium was replaced by the medium containing the GMO-MENs at a differential concentration of 0–100 μg/ml per well and the cells were incubated for another 24-h period. Then, the cell medium was replaced with a fresh one and washed with the PBS buffer and cell viability assay was performed by adding 50 μl per well of XTT-activated solution from the XTT test kit supplied by ATCC and incubate for 4 h at 37°C. The experiments were performed in triplicates.

11.4.3.7 Heat Dissipation due to Field Treatment with MENs In this experiment, the temperature was measured locally via IR camera FLIR-i3 on the surface of both cancer (SKOV-3) and healthy (HOMEC) ovarian cells before and after a field treatment. The experimental error of the setup was approximately ±2°C of the infrared camera. The magnetic field of 30 Oe was applied for a 24-h period. No significant heat dissipation was observed as a result of the field treatment. The negligible heat dissipation (compared to the conventional method) is a consequence of the intrinsic nature of the ME coupling which resulted in the relatively high high-efficacy control of intrinsic electric fields by external magnetic fields.

11.4.3.8 Universal Applicability: MENs-Triggered Drug Uptake in MDR Cell MES-SA/DX5 To demonstrate the applicability of the new nanotechnology to other cancers, we conducted a parallel study on a well-known multi-drug-resistant cell line MES-SA/DX5 [24]. For comparison, the following four different drug–delivery system combinations were studied: (i) no drug, (ii) free Flutax-2, (iii) Flutax-2-GMO-MENs with no field, and (iv) Flutax-2-GMO-MENs with 30 Oe field.

11.4.4 Discussion

The results of the previous experiments confirmed our hypothesis that MENs loaded with the drug PTX could serve as high-specificity remotely controlled (via magnetic fields) delivery nanosystems to treat EOC. This function is believed to be achieved due to the localized electroporation effects induced by the MENs in the vicinity of the cancer cell membranes when exposed to an external magnetic field. This effect at the nanoscale, was described by a new terminology, "NANO-ELECTROPORATION." The main hypothesis was implemented in the experiment as a twofold process, each distinctly highlighting the two core field-dependent processes defined by two critical fields: (i) the threshold field, H_{th}, for MENs to penetrate through the cancer cell membrane to deliver the drug into the cell cytosol (by means of the field-induced localized nano-electroporation effect in the vicinity of MENs) and (ii) the release field, H_r, that triggers unloading of the drug after the drug-loaded MENs penetrated into the cell. Both the typical HER-2 antibody chemistry and the new physical mechanism that relied on the significant difference in the threshold electric field between the healthy

and cancer cells defined the specificity of the drug to the cancer cells. This threshold field was measured to be of the order of 30 Oe and above 200 Oe for the SKOV-3 and HOMEC cells, respectively. The experiment also proved the increased eradication of the cancer cells via this remote magnetic field-triggered electric field-defined specificity to the cancer cells. Compared to the traditional antibody-mediated targeting (fig. 4 [12]), the percentage of the cell-penetrated drug was increased by a factor of five in this case. Moreover, the result showed that after the drug was efficiently transferred through the tumor cell membrane by the field-controlled MEN-initiated nanoelectroporation, within a 24-h period of a low-energy 30-Oe treatment (figs. 11 and 12) [17], the majority of the cancer cells were eradicated without affecting the healthy cells in their vicinity.

The value of the release field, H_r, was chosen to be higher than the value of the threshold field to penetrate through the membrane, H_{th}, for the cell of the same type. This was done to achieve adequately high efficacy of the drug delivery. The release field is defined by the binding force between the MEN and the drug, while the penetration threshold field, H_{th}, is mostly determined by the electric properties of the cell membrane that lead to the localized electroporation effects. The release field could be controlled by the proper selection of the intermediate layer between the drug and the MEN. In addition, the release field depended on the field treatment duration and the frequency of the a.c. field, as shown in figure 8 [17]. For example, as shown in figure 8A [17], the spectrophotometry measurements of the absorbance at 230 nm (for PTX) indicated that only 1 min of field treatment at a 66-Oe d.c. magnetic field was sufficient to release over 95% of the drug. As shown in figure 8B [17], the same release efficacy (of over 95%) could be achieved also at an 1000-Hz a.c. field at a smaller field strength of 44 Oe in 2 h of treatment. This complex dependence can be explained by the fact that the external field effectively reduces the energy barrier that holds together the MEN and the drug, while an increase of the treatment duration increases the temperature-induced probability to overcome the barrier or breaking the bond. Here, it can be mentioned that although using a.c. external magnetic fields could indeed add another knob to control the targeted delivery, in this study to focus on the proof of the main hypothesis, as illustrated in figure 7 [17], we followed the d.c. field scenario. The d.c.-field-controlled drug release kinetics was confirmed also through AFM, FTIR, and mass spectrometry, and XRD pattern studies [24]. As confirmed by IR measurements of the cellular surface temperature, the MENs' field action didn't trigger any significant temperature changes in the field and frequency range under study. This is in agreement with the fact that the MENs-induced delivery is a relatively energy-efficient process (because of the strong intrinsic ME coupling), which causes only negligible heat dissipation.

As for the penetration threshold field, H_{th}, we found that for the cancer cell membrane the value was of the order of 30 Oe, that is, less than the d.c. release field (for the default MEN carriers coated with GMO) of 60 Oe. Again, according to the main hypothesis, to ensure the specificity to the cancer cells, it is important to maintain the remote field above the release value for the tumor cells but lower than the threshold value for the healthy cells. Indeed, at a 30-Oe external d.c. field, the drug couldn't penetrate through the healthy cell membrane for the 24-h treatment duration, which

confirmed that the threshold field for the drug-loaded MENs to penetrate into the healthy cell exceeded 30 Oe during the entire treatment (fig. 11 [17]). Specifically, the GMO-MENs field-treated HOMEC cells showed negligible drug intake per 1 mg of the cellular protein content. The value was 0.18 ± 0.07, 0.30 ± 0.04, and $0.55 \pm 0.16\%$ for the field strength of 5, 15, and 30 Oe, respectively. On the contrary, SKOV-3 cells showed significantly higher values of the drug intake, which was 1.50 ± 0.41, 2.36 ± 0.48, and $10.41 \pm 1.54\%$ for the field strength of 5, 15, and 30 Oe, respectively. It can be noted that after a 24-h 30-Oe field treatment by GMO-MENs, approximately 95 and 34% of HOMEC and SKOV-3 cells, respectively, remained viable. When the treatment was extended to 36 h, the percentage of viable cells fell to approximately 85 and 10% for HOMEC and SKOV-3 cells, respectively. These results indicate that further field and frequency optimization could be used to perfect the treatment results.

In addition, the cytotoxicity measurements of the standard XTT assay performed on SKOV-3 cells at different concentrations of MENs showed no significant toxicity even at the highest nanoparticles concentration value of 100 μg/ml. The chart shows the results of XTT assay performed on SKOV-3 cells at different concentrations of GMO-MENs.

Finally, the parallel study on MDR cell MES-SA/DX5 proved the applicability of the new nanotechnology to other cancers. It might be worth noting that due to the overexpressed transmembrane proteins, for example, P-glycoprotein, this cell type is known to be relatively impenetrable for many popular chemotherapy drugs, which makes the finding even more significant [25, 26].

11.5 EXPERIMENT 3: MAGNETOELECTRIC "SPIN" ON STIMULATING THE BRAIN

A neural network can be considered as a complex bioelectric circuit made of many neurons connected through chemical and electrical synapses formed between axons and dendrites [27–29]. The signaling in the network is electric field-driven and based on a highly collective system of electric charges, neurotransmitters and action potentials. The ability to remotely incite specific neuronal excitations deep in the brain with the purpose of artificially stimulating selective regions of the network remains an important open question in neural engineering [30]. Furthermore, the ability to control the CNS at micro- or even nanoscale could provide unprecedented control of specific functions and enable highly personalized "pinpoint" treatment of neurodegenerative diseases such as Parkinson's disease, essential tremor, epilepsy, and others [31, 32]. The current brain stimulation technology is operated at macroscale and often relies on highly invasive direct-contact-electrode techniques such as deep brain stimulation (DBS) [33–35]; it can be noted that DBS is one of only a few neurosurgical methods allowed for blinded studies. There are also noninvasive brain stimulation methods; these include repetitive transcranial magnetic stimulation (rTMS) [36, 37] and transcranial d.c. stimulation [38, 39]. Though rTMS and transcranial d.c. stimulation indeed represent major technological advances in noninvasive brain

stimulation, the depth and locality of focusing are strongly limited in both methods [40–42]. In rTMS, relatively high intensity magnetic fields (on the order of 10,000 Oe) are required to stimulate regions deep in the brain; however, high intensity magnetic fields, especially in the a.c. mode, may lead to excessive energy dissipation or other destructive side effects [43]. The required high external magnetic field can be explained by the relatively weak coupling between the magnetic field and the local electric currents in the neural system. With that already said, one can identify the following engineering bottleneck. On the one hand, using electric fields one can achieve adequate brain stimulation; however, the need to establish direct contact with individual neurons makes the electric field stimulation highly invasive. It is hard to see how such a physical direct contact with each of the 75 billion neurons in the brain can be achieved. On the other hand, magnetic fields can penetrate through the brain without being significantly distorted by the local microenvironment and therefore can be used for externally controlled stimulation; however, inadequately weak coupling between external magnetic fields and intrinsic neural activity-induced electric currents makes such stimulation relatively inefficient and consequently inadequate for being used for local stimulation deep in the brain.

One potential solution for enabling high-efficacy remote control of the neural activity deep in the brain would be to use conventional MNs to locally amplify the magnetic field and thus enhance the effective coupling to local electric currents [12, 44]. An even more dramatic solution could be achieved by using a novel class of functional nanoparticles known as MENs [45–47]. Under an equivalent magnetic field exposure, MENs can not only amplify the local magnetic field but also generate an additional local electric field. The field is generated due to the nonzero ME effect, which originates from an intrinsic coupling between electric and magnetic fields in MENs. As a result, when administrated in the brain, MENs can serve as nanoscale sites which, when exposed to a relatively low external magnetic field (in the range of 100–1000 Oe), generate local electric fields (on the order of 1000 V/m or higher) to provide direct external access to the internal (bioelectric) neural circuits. In other words, MENs enable a unique way to combine the advantages of both the high efficacy of the electric fields and the external-control capability of the magnetic fields, therefore opening a novel pathway to control the brain. To the best of our knowledge, this chapter for the first time presents results of an animal study to demonstrate the use of toxicity-free MENs to stimulate the brain via application of low external magnetic fields.

11.5.1 Methods

11.5.1.1 Fabrication of Toxicity-Free MENs Details on the fabrication of toxicity-free MENs were described in our recently published comprehensive studies [48, 49]. The basic structure of $CoFe_2O_4$–$BaTiO_3$ core-shell 30-nm MENs was synthesized according to the following steps: 0.058 g of $Co(NO_3)_2{\cdot}6H_2O$ and 0.16 g of $Fe(NO_3)_3{\cdot}9H_2O$ were dissolved in 15 ml of deionized water; 5 ml of aqueous solution containing 0.9 g of sodium borohydride and 0.2 g of polyvinylpyrrolidone was added at 120°C for 12 h to obtain $CoFe_2O_4$ nanoparticles; $BaTiO_3$ precursor solution was

prepared by adding 30 ml of deionized water containing 0.029 g of $BaCO_3$ and 0.1 g of citric acid to 30 ml ethanolic solution containing 1 g of citric acid and 0.048 ml of titanium (IV) isopropoxide; as-prepared $CoFe_2O_4$ nanoparticles (0.1 g) were added to the 60 ml of $BaTiO_3$ precursor solution and sonicated for 120 min; the resulted dispersed nanoparticles were dried on a hot plate at 60°C for 12 h while stirring at 200 rpm; the obtained powder was heated at 780°C for 5 h in a boxfurnace and cooled at 52°C/min to obtain 30 nm-sized $CoFe_2O_4$–$BaTiO_3$ core-shell MENs. To eliminate toxicity at least in *in vitro* studies, the nanoparticles were surface functionalized by a 2-nm thick coating of GMO according to the following steps: GMO-MENs were prepared by incubating 0.1 mg of GMO with 5 mg of MENs in 5 ml of PBS (pH 7.4) buffer for 12 h to achieve uniform surface modification and the solution was slowly agitated during incubation; the solution was centrifuged at 20,000 rpm for 20 min at 10°C to remove excess GMO; the obtained pellet was resuspended in ethyl acetate : acetone (70:30) solution and recentrifuged three times to obtain GMO-MENs; surface-modified MENs were lyophilized and stored at 4°C until further use; the particle size distribution was measured by a Zetasizer Nano series that uses the standard dynamic light scattering approach, and the measured zeta potentials for nonfunctionalized and functionalized (with GMO) MENs were -45 ± 1.72 and -41.6 ± 0.26 mV, respectively.

11.5.1.2 An* In Vitro *Cytotoxicity of MENs An *in vitro* cytotoxicity study was performed using XTT assay (sodium 2,3-bis(2-methoxy-4-nitro-5-sulfophenyl) 5-[(phenylamino)-carbonyl]-2*H*-tetrazolium inner salt) according to the manufacturer's protocol (ATCC), as previously discussed [49, 50]. The assay was performed on human astrocyte cells and peripheral blood mononuclear cells for neuronal and peripheral cytotoxicity, respectively. The results of the XTT assay (performed in triplicates, $n = 3$) indicated no significant toxicity on both human astrocyte and peripheral blood mononuclear cells, respectively, in the GMO–MEN concentration range from 0 to 200 μg/ml.

11.5.1.3 Animal Studies: Electroencephalography Setup To directly track the electric field response in the brain under different experimental conditions of this study, a two-channel electroencephalography (EEG) headmount was surgically implanted into the head of an imprinting control region (ICR) mouse (~2-month-old females). As described later in more detail, four solid electrodes, two for each channel, were surgically attached to the skull via four nonmagnetic stainless steel screws to physically touch the brain in four different points over the cerebral cortex. This setup allowed to directly measure local electric signals in the brain's neural network in response to external-magnetic-field-activated nanoparticles. During the measurements, mice were kept under general anesthesia using inhalatory isoflurane (~1.5%) through a stereotaxic nosecone. The nanoparticles were introduced into the bloodstream via IV injection through the tail vein. The concentration of the nanoparticle saline solution was approximately 200 μg/ml. The amount of the solution injected into the tail vein was approximately 0.05 ml (~10 μg of nanoparticles). After the injection, to transfer the nanoparticles from the bloodstream into the brain across

BBB, a magnetic field gradient of approximately 3000 Oe/cm was applied for approximately 30 min using a permanent neodymium magnet placed over the top part of the head for approximately 30 min. Because the main focus of this study was to understand the novel properties of MENs in the neural microenvironment in the brain, no significant emphasis was made on transferring all the IV-administrated nanoparticles across BBB. Generally, it is believed that approximately 10% of IV-administrated nanoparticles transfer into the brain [50]. In this particular case, this estimate might be conservative considering a "magnetic-field-gradient pull" was used to force the nanoparticles to cross BBB. After the "smart" nanoparticles were delivered in the brain, the response of the two-channel EEG headmount to local neural electric signals due to the nanoparticles' magnetic (with MNs) or magneto-electric (with MENs) response was studied via application of a low-energy a.c. magnetic field with a magnitude of 100 Oe and a frequency in a range from 0 to 20 Hz. The a.c. field was generated by a 1-inch electromagnetic multiturn coil that was placed approximately 0.5 cm away from the head. The field profile generated by the standard coil setup can be easily calculated and therefore was well predictable and controlled. To avoid any potential interference of the fields due to residual nanoparticles, a separate mouse was used for each nanoparticle type and dosage.

11.5.1.4 Animal Studies: Surgery The current *in vivo* studies were performed in accordance with the guidelines of Institutional Animal Care and Use Committee (IACUC) # FIU: 13–045. The ICR mice under study were purchased at The Jackson Laboratory for Genomic Medicine. Mice were sedated in a 5% isoflurane environment in an induction chamber before they were placed on a stereotaxic apparatus for further procedures under general anesthesia with a 2% isoflurane flow through a gas mask. A surgical incision was made on the head to expose the periosteum. A two-channel EEG headmount was affixed onto the periosteum using a surgical glue (cyanoacrylate) and four 3-mm stainless steel screws which also served as EEG electrodes. The two channels were separated by approximately 5 mm. The headmount was placed approximately 3 mm anterior to bregma, so that all the four screws would rest on the cerebral cortex region of the mouse brain. After the electric contact between the headmount and the EEG electrodes was re-enhanced via application of silver epoxy, the implant was further secured via a dental acrylic. As a postoperative care, every 8 h for 3 days, mice were administered with analgesics (Buprenex, 0.15 mg/kg) through subcutaneous injection and antibiotics (Baytril, 2.27 mg/ml) through drinking water. Mice were given 2 weeks for recovery before conducting further experiments. Upon completion of the study, the mice were sacrificed by CO_2 hypoxia followed by cervical dislocation according to the American Veterinary Medical Association (AVMA) guidelines.

11.5.1.5 Cryosectioning of Brain Tissue Whole brains were dissected and immersion-fixed in 4% paraformaldehyde (wt/vol; dissolved in PBS solution) at 4°C overnight with gentle shaking. Then, the brains were washed three times in PBS solution, cryoprotected with PBS-buffered 30% sucrose and embedded in optimum

cutting temperature medium (Tissue-Tek). Serial 10-μm-thick sections along the sagittal plane of the right and left cerebral hemispheres were cut using a Leica CM3050 Cryostat. AFM imaging was performed directly on the cryosectioned samples.

11.5.1.6 Various Microscopy Sample Preparations *Scanning electron microscopy (SEM) sample preparation*—Cryosectioned mouse brain tissue samples were fixed by a 3% glutaraldehyde buffer solution (0.1 M PBS) at room temperature for 2 h. After the samples were washed three times with a 0.1 M PBS (pH 7.2), they were immersed in a 1% osmium tetroxide solution (250 mg of OsO_4 in 25 ml of 0.1 M PBS, pH 7.2) for 2 h at room temperature in a dark environment, followed by washing three times with a 0.1 M PBS, with a pH of 7.2. After washing, dehydration was induced by graded ethanol solution in water for 5 min for the following concentration sequence: 30, 50, 70, 80, 90, 95, and 100%. Then, the samples were dried in a graded HMDS ethanol solution of 30 and 50% for 20 min and 100% for approximately 14 h. Thus prepared samples were mounted on an SEM specimen stub with a carbon tape for imaging.

Vibrating sample magnetometry—A room-temperature Lakeshore vibrating sample magnetometer (VSM) with a 3-T magnetic field sweep was used to measure key magnetic properties of nanoparticles under study including the magnetization saturation and the magnetic coercivity.

Scanning probe microscopy—Scanning probe microscope multimode was used to measure AFM and magnetic force microscopy (MFM) images of nanoparticles under study. The MFM images were obtained with a CoCr-based hard magnetic nanoprobe at a scan height of 10 nm.

Transmission electron microscopy—Phillips CM-200 200 kV TEM with energy-dispersive spectroscopy (EDS) option was used to obtain TEM images and EDS profiles.

11.5.2 Results

A typical TEM image of the synthesized 30-nm MENs, with a clearly visible core-shell structure, is shown in figure 13A [17]. The core and shell were made of the magnetic-moment-enhancing $CoFe_2O_4$ ferromagnetic spinel and magnetoelectricity-inducing $BaTiO_3$ perovskite nanostructures, respectively. The composition was confirmed through energy-dispersive spectroscopy. These nanostructures have an ME coefficient that can be varied from 1 to 100 mV/(cm Oe) through the substitution of transition metals by other elements from the same series. A typical room-temperature magnetic hysteresis loop, which was measured via VSM, is shown in figure 13B [17]. The curve indicates a ferromagnetic nature of the magnetic component. The saturation magnetization and the magnetic coercivity of these nanoparticles were on the order of 1 emu/g and 100 Oe, respectively. Figure 13C [17] shows an MFM image of MENs subjected to an external electric field of approximately 100 V/cm along the *X*-direction. The image, obtained with a low moment probe, shows an electric-field-induced dipole moment in the same orientation. Typical SEM images of brain sections in a sagittal plane for two mice, including a sample with MENs and

a sample without any MENs (control), respectively, are shown in figure 14A [49–51]. A pair of AFM and MFM images of a "zoomed-in" region of the brain section with MENs are shown in figure 14B [17]. The EEG waveforms and their respective frequency spectra (Fourier transforms) for the response to a 100-Oe external magnetic field at frequencies of 0, 5, 10, 15, and 20 Hz are shown in figure 15 [17]. The frequency range was limited by the available electronics. The resulting modulation by the a.c. magnetic field can be clearly observed. The modulated periodic signal on the order of 1 mV is comparable to the magnitude of the unperturbed original signal in the brain. As seen in the frequency response, the new modulated electric field component is at the same frequency as the external a.c. magnetic field applied by the coil setup, as described earlier. The two channels show comparable signals.

Thus determined intrinsic electric field modulation amplitude was measured as a function of the external a.c. magnetic field amplitude, as shown in figure 16 [17]. The dependence is saturated at approximately 100 Oe field on the order of the coercivity field of the MEN system under study. To confirm that the modulated effect indeed resulted from the presence of MENs, the following control measurements were performed. In the first control measurement, EEG waveforms were recorded without introducing any nanoparticles in the brain. The typical EEG waveforms at five values of the a.c. frequency, 0, 5, 10, 15, and 20 Hz, respectively, are shown in Supplementary figure 15 (left) [17]. The frequency spectrum (Fourier Transform) of the signal for each a.c. frequency response is shown in Supplementary figure 15 (right) [17]. Note that the waveforms were not visibly affected by application of the a.c. magnetic field in the frequency (<30 Hz) and amplitude ranges (<300 Oe) under study. In the second control measurement, MENs were substituted with the same amount of conventional MNs. As conventional MNs, 30-nm $CoFe_2O_4$ nanoparticles were injected in the brain at the equivalent dose of 10 μg. The saturation magnetization of these nanoparticles was on the order of 100 emu/g (via VSM measurements), compared with approximately 1 emu/g for MENs. The typical EEG waveforms from the two EEG channels and their frequency spectra for the response to a 100-Oe external magnetic field at frequencies of 0, 5, 10, 15, and 20 Hz are shown in Supplementary figure 16 (left) and (right) [17], respectively. Note that like the first control measurement (without any nanoparticles in the brain), the waveforms are barely affected by the application of an a.c. magnetic field. Here, one should mention that the experiment could not completely exclude the effects due to MNs; it just shows that the sensitivity of the current setup (~50 μV) was not sufficient for detecting the MN-induced signal under the current experimental conditions.

11.5.3 Discussion

The main goal of this study was to demonstrate through *in vivo* studies on ICR mice that MENs, when administrated in the brain, could enable externally controlled non-invasive DBS [52]. In the main experimental setup, IV-injected nanoparticles were forced via application of a magnetic field gradient of approximately 3000 Oe/cm to enter the brain by crossing BBB (fig. 15 [17]). Like conventional MNs, MENs have a nonzero magnetic moment, and therefore could be controllably transferred across

the BBB via application of a remote magnetic field gradient force, confirmed through two types of imaging, SEM and AFM/MFM, respectively, to observe the presence of MENs in brain sections (fig. 14 [17]). The MENs' coercivity on the order of 100 Oe (fig. 13B [17]) was smaller than the stray field generated by a typical CoCr-based MFM probe, and therefore during imaging the nanoparticles were always magnetized in the same orientation as the nanoprobe's magnetic moment. The dark color for the nanoparticles in the MFM image reflects the attractive force between the nanoparticles and MFM nanoprobe, which is in agreement with their relative orientations. Note that the particle distribution was relatively inhomogeneous, as we would expect. From the signal-to-noise perspective, to have a substantial portion of the brain to be sensitive to the stimulating effect, it would be preferred to have as many nanoparticles per neuron as possible without destructing the operation of the brain. A dose of 10 μg of MENs (equivalent to ~20 billion nanoparticles, assuming the atomic density of a 30-nm nanoparticle to be on the order of 5 g/cc), was administrated in the bloodstream through an IV injection in a mouse tail vein. Considering 75 million neurons per brain, on average, there were hundreds of particles per neuron. It should be noted that the current estimate is very approximate. In the future, a more comprehensive *in vivo* study must be conducted to understand the detailed metabolism of the nanoparticles. A special EDS-based analysis of the nanoparticle distribution in different organs has been developed to study the biodistribution of MENs of different compositions, sizes, and shapes in different organs in different time intervals after the injections. This study will be discussed in a separate chapter in the near future. An example of the biodistribution analysis of nanoparticles in a kidney tissue is shown in Supplementary figure 17 [17]. Today, iron oxide MNs are used as image contrast enhancement agents in magnetic resonance imaging [52, 53]. It can be assumed that MENs undergo a similar metabolism as these conventional nanoparticles do. In addition, the current study cannot provide the information on the distribution of the nanoparticles in different cells in different brain regions such as, cerebral vascular cells, lymphocytes, astrocytes, and others. In a first-order approximation, the distribution can be assumed to be similar to the distribution of long-circulating PEGylated cyanoacrylate nanoparticles [51]. Finally, it is worth noting that the novel nanotechnology allows using external magnetic fields to determine the location of the stimulation regions deep in the brain. One potential approach to accomplish the spatial selectivity might be the implementation of multiple coherent near-field antennas that operate in a hologram-like regime.

As noted earlier, MENs, unlike MNs, have an entirely new property due to the ME coupling. Because of its intrinsic nature, the coupling is relatively energy efficient. For example, considering the value for the ME coefficient, approximately 100 mV(cm Oe), according to a trivial isotropic expression for the ME effect, $\Delta P = \alpha H$, where P and H stand for the induced electric dipole field and the external magnetic field, respectively, an electric field on the order of 100 V/m could be generated a few nanometers away from a MEN merely by applying a weak magnetic field of 100 Oe. This would be a conservative estimate, because no contribution of the edge effects due to the deviation of the nanoparticle's shape from an elliptical configuration was taken into account. It can be noted that the magnitude of this induced local electric field is

comparable to that of the electric field due to the normal neural activity in the brain, for example, displayed as action potentials. Previously, using an example of patients with Parkinson's disease, we have theoretically shown that indeed the ME coupling in intravenously injected MENs could be used to stimulate the brain via remote application of a relatively low magnetic field (~100 Oe) in the near d.c. frequency range (~< 100 Hz) to the same level achieved via the highly invasive conventional DBS [4]. In this case, each MEN in the brain serves as a highly localized high-efficacy converter of a weak external magnetic field into a high internal electric field; consequently, the functions in the brain could ideally be controlled at the subneuronal level provided the nanoparticle dosage and spatial distribution are selected so that at least a few nanoparticles are present per each neuron. Still, an important milestone of proving the concept experimentally and particularly through *in vivo* measurements was achieved only in the current study. The two EEG channels, which were surgically attached to the head, could directly measure the local electric signals in two regions of the brain. As predicted, using an externally applied a.c. magnetic field with an amplitude of only 100 Oe and in a frequency range from 0 to above 20 Hz, we could significantly modulate the two EEG waveforms with the frequency of the external a.c. magnetic field (fig. 15 [17]). The amplitude of the modulated sinusoidal signal at four frequency values, 5, 10, 15, and 20 Hz, respectively, was comparable to the magnitude of the original nonmodulated signal at the same locations. The fact that the modulation amplitude saturated at an a.c. external magnetic field of 100 Oe (fig. 16 [17]), which was comparable to the coercivity of MENs, was in agreement with the assumption that the electric field stimulation was triggered by the signal coming from MENs due to an externally generated magnetic field. At the same time, the fact that the a.c.-induced modulation signal disappeared after the mouse was euthanized indicates that the modulation signal indeed originated from the active neural network and was not caused by an inductive coupling between MENs and the EEG probes. To further ensure that this modulated signal was caused by the presence of MENs in the brain and did not originate from the electrical setup contacts/connections or other sources, we conducted equivalent control measurements without any injected nanoparticles.

No detectable frequency modulation was observed in this case. Here, it should be restated that to avoid any signal interference due to any residual nanoparticles in the brain, each new set of measurement conditions was tested on a new mouse. Finally, to compare MENs to the conventional MNs that do not display the ME effect, another control study was conducted; a similar dose of MNs was administrated according to the aforementioned standard procedures. These 30-nm MNs were made of a high-moment $CoFe_2O_4$ composition, with a coercivity of 100 Oe and a saturation magnetization of 100 emu/g. The applied a.c. field of 100 Oe was comparable to the coercivity field for MNs. As a result, because of the relatively high saturation magnetization value, MNs in their proximity were expected to regenerate a magnetic field on the order of 1000 Oe, in other words, at least an order of magnitude higher than the externally applied field. It is known that such high magnetic fields can indeed stimulate the brain. For example, the already discussed rTMS approach is based on such high fields. Nevertheless, compared with the modulation

signal due to MENs, the signal due to MNs was below the detection level of approximately 50 μV. This observation is in agreement with the significantly stronger magnetic-to-electric-field coupling in the case of MENs due to the intrinsic nature of this coupling. Strikingly, in this comparison between MENs and MNs, the saturation magnetization of MENs was two orders of magnitude smaller than that of MNs; the latter indicates that the effect of the MEN-generated local electric field overshadows the effect of the MN-regenerated local magnetic field. Unlike stimulation with MNs or rTMS, stimulation with MENs did not require a high external magnetic field on the order of 104 Oe and instead originated from a significantly enhanced local high electric field only a few nanometers away from the nanoparticles that were exposed to a 100 Oe external magnetic field. Apparently, because there are hundreds of nanoparticles per neuron, the MENs' contribution is adequately strong to trigger the observed significant neural stimulation. Though both MNs and MENs respond to an external magnetic field, they respond differently. When exposed to a magnetic field, MNs can locally regenerate an even stronger magnetic field. The field amplification factor, β, is proportional to the ratio of the saturation magnetization, M_s, and the magnetic anisotropy field, H_K, in other words, β is approximately $4\pi M_s/H_K$. For example, for typical iron oxide-based nanoparticles with $4\pi M_s$ on the order of 1000 emu/cc and H_K on the order of 100 Oe, the local amplification factor would be on the order of 10, and therefore the locally amplified magnetic field would be on the order of 1000 Oe. In turn, the higher local magnetic field implies the stronger coupling to intrinsic electric currents due to the local neural activity. As for MENs, under the same magnetic field exposure, not only do they modify the local magnetic field, but they also generate a local electric field.

Therefore, because the neural system is inherently electric field driven, these new "smart" nanoparticles can provide substantially stronger coupling to the system. In fact, typical MENs have a substantially smaller value of the saturation magnetization (often, a couple of orders of magnitude smaller) compared with that of typical MNs; thus, in case of MENs, the magnetically induced electric field action can substantially overshadow the purely magnetic field action. In summary, while in theory both MNs and MENs could solve the current problem of the weak coupling between external magnetic fields and CNS's internal electric fields, MENs could provide a more intricate coupling to CNS compared with that by MNs. Because the two nanoparticle systems under study, in other words, MNs and MENs, have a nonzero magnetic moment, they can ideally be detected via existing magnetic imaging techniques such as the conventional MRI or the state-of-the-art MN imaging (MNI) [54–56]. If the concentration of nanoparticles is high enough to ensure that at least a few of them are present in each neuron, the selected brain region would be stimulated without a need for the earlier conventional invasive contact-based DBS or relatively inefficient high-field TMS procedures when exposed to a local magnetic field. In other words, by introducing the MENs in the brain, we effectively create a new "nanoenvironment" in which intrinsic electric signals in CNS are robustly coupled to an external magnetic field at the subneuronal level.

Such an energy-efficient external connection to the brain could potentially open an unprecedented pathway to control selective brain regions; magnetic imaging

techniques such as MNI or MRI could be used for both 3D imaging and selective local stimulation on demand by localizing fields via special magnetic coils. (Here, it could be mentioned that MNI might be more suitable for being used with MENs because of the potential for superior sensitivity and real-time resolution capability.) The basic steps of the main concept are illustrated in schematics in figure 17 [17]. The current study was focused on the concept demonstration. To further underscore the significance of this study, it is worth noting that introduction of MENs in the brain could open a pathway for "writing" and "reading back" an electric field map of the brain at a subneuronal level. Such 3D electric field mapping via MENs can be used for both remote stimulation of specific functions and "reading back" the electric field information with a subneuronal precision, provided that there is an adequate magnetic imaging technique available. For example, one important future application of the unprecedented capability would be to use MENs in conjunction with existing magnetic imaging approaches, for example, MRI or MNI (suitable for real-time monitoring), to enable "pinpoint" diagnostics and therapy. While magnetic imaging using MENs as contrast agents would be modulated with local electric fields due to the neural activity and therefore would provide a real-time electric field map of the brain, a focused low-energy a.c. magnetic field could be used to stimulate/treat any specific region deep in the brain in a "pinpoint" fashion. Theoretically, according to the principle of reciprocity, proving the validity of one of the two ways, "writing" or "reading back" information in the brain, respectively, can lay ground for proving the validity of the other [57]. The focus of this particular study was to prove that the brain could be "written" (stimulated) through an external field. Specifically, we demonstrated that with MENs, because of their strong ME coupling, a portable battery-driven and relatively small-sized coil could generate an external magnetic field that is strong enough to stimulate intrinsic electric signals in the CNS deep in the brain. Finally, it could be noted that the MEN's parameters are not limited to the $CoFe_2O_4$–$BaTiO_3$ composition and the 30-nm diameter, as used in the current study. It is likely that the concept could be further improved through using different materials, such as, biodegradable compositions, and even smaller-sized nanoparticles.

11.6 BIOCERAMICS: BONE REGENERATION AND MNS

Another important application of MNs can be found in the tissue regeneration. In particular, using bioceramic materials in bone and tissue regeneration is an interesting proposition. These bioceramic materials are capable of acting as bridges and giving support much like the bones or tendons themselves. This property of their acting as scaffolds can be used for repairing and maintaining damaged tissues in the same manner as a cast helps heal broken bones. Some of these ceramic materials like calcium phosphate glass ceramics can be used to deliver hormones, drugs, and nucleic acids in a controlled manner, which help in regrowth of tissues.

In order to prevent formation of bacterial biofilms and postsurgical infections many times, these bioceramic materials are coated with antibiotics. Mesoporous bioactive glass has outstanding drug-carrying characteristics. They have been found

to be osteoconductive, osteoproductive, and osteoinductive, making them perfect candidates for the job.

With the development of MNs, they can be used as thermoseeds. Under an alternating magnetic field, they can be used to generate heat in the neighboring regions thus causing hyperthermia of cancer.

In the following subsections, we will be discussing how bioceramics also exhibit selective targeting properties, to be followed by the section on MNs that talks about the use of MNs in hyperthermic treatment of cancer and a potential candidate for targeted delivery. Last, we talk about the integration of MNs with bioceramic silica to form mesophoric structures.

11.6.1 Selective Mediators for Cancer Targeting

Nanomaterials especially magneto nanoparticles exhibit controllability, which makes them a desired candidate as drug carriers for targeted delivery. Hence, it is not difficult to believe that nanoparticles are the most promising field in biomedical applications. The needs of the medical field are being fulfilled by advances in the nanotechnology. For example, research conducted by Tsai and coworkers have used mesoporous silica nanoparticles for selectively targeting breast cancer cells by specific monoclonal antibodies [58].

Research has shown that bioceramic polymers and polymeric nanofibers can exert control in the release of therapeutic drugs as compared to conventional drug delivery systems [59, 60]. By appropriately modifying the surface area of the polymer or the composition of bioceramic component, the rate of release can be controlled. In the following subsections, we will be discussing about MNs and bioceramics.

11.6.2 Magnetic Micro- and Nanostructures

Hyperthermia is a process of exposing the body cells to slightly elevated temperatures to either kill cancer cells or make them prone to anticancer drugs. Magnetic microspheres, because of their property to heat up in presence of magnetic field, can be used to trigger hyperthermia in cancerous tissues. In Ref. [61], they talk about systemic delivery of these magnetic particles and after they get accumulated at the site of the tumor, heat can be generated by these magnetic particles by applying an alternating magnetic field. The alternating magnetic field induces hysteresis losses, which manifests into heat. We propose a further addition to this application. We know that MNs can also be controlled by an external magnetic field to target at a particular location within the body, so we can use this property to our advantage also, instead of using systemic delivery we can exert magnetic control for better selectivity.

Thus we see that hyperthermia is not the only way MNs can be benign, they can assist in more than one ways like selective targeting also, not to forget MRI and gene transfer. In Section 11.6.3, we will be discussing about the targeted release of drugs.

11.6.3 Magnetic Drug Delivery System

As we saw in the earlier section of the chapter which talks about MENs, most of the therapeutics involving MNs have a core-shell type of configuration in which various composites like silica, non-polymeric organic stabilizers, and polymeric surfactants may coat MNs. In Ref. [62], they come up with the first ever core-shell magnetic silica dispersions. The Stöbler method was used [63] for synthesizing this product. In this process, silica particles grow in an alkaline mixture of ethanol and water. This coating of silica on magnetic core such as the ones described in the earlier portion of this chapter increases the stability of the magnetic particles and also makes them biocompatible. These nanoparticles can act as carriers of therapeutic substances in the silica shell for targeted drug delivery, or they can attach fluorophores for cell marking [64]. Another way of fastening the drugs is to encapsulate the drug between the layer of silica shell and the magnetic core of the nanoparticle [65]. This design helps in controlling the release of the drug by an external magnetic field.

One of the important characteristics of the silica shell is that it provides mechanical and chemical stability to the magnetic core as well as the drugs incorporated within. For controlled release of the drug and targeted delivery, it is important they are adsorbed on the surface. The creation of mesoporous materials using surfactant structure directing agents ensures that the adsorption process takes place smoothly. Another important thing to be noted is that excessive amounts of magnetic cores can be injected in a single step, so that the targeted area can reach temperatures required for hyperthermia to take place under the influence of alternating magnetic fields. All these features allow merging hyperthermia treatment and chemotherapy to fight against solid tumors. The mesoporous ordering can be varied by changing the ratio of surfactant to silica and the magnetic properties can be altered by varying the number of encapsulated MNs. In Section 11.6.4, we will be talking more about magnetic mesoporous silica carriers and what are the different ways to control the drug delivery.

11.6.4 Magnetic Mesoporous Silica Carriers

Magnetic mesoporous microspheres are MNs having a network of pores with diameters comparable to the dimensions of the drug that they are intended to carry. Their pore size distribution is tunable. The pore size can be altered in the self-assembly process by using different surfactant molecules. In an experiment mentioned in Ref. [66], by employing an aerosol-assisted method as illustrated in figure 18 [66], and using non-ionic Pluronic P123 as structure directing agent, matrices with pore sizes of 5.6 nm were produced. The crystallinity as well as the superparamagnetic behavior of the MNs was preserved even after the coating of silica was applied. Since the magnetic properties of the core MNs are preserved even after the pyrolysis process as was evident from earlier, they still possess the ability to be controlled by an external magnetic field. The silica microparticles present a set of hexagonal array of pores. The specific surface area and pore volume of the hexagonal arrays thus synthesized is more comparatively. Also, it was found to carry and release the drug

as expected. In the case study presented in Ref. [67], two different views of release in two different regions were presented. In one case during the first hours of the assessment, a 55% drug delivery was recorded for an initial burst effect. In the second case, a more controlled kinetics was involved. This pattern is a fit for cases that require high initial doses, and successively maintaining medication levels.

As mentioned earlier, these mesoporic silica particles are characterized by stable mesoporous structure, high specific surface area, and adjustable pore sizes as well as pore volume. All these properties make them excellent candidates for the controlled release of drugs. Along with that, they also allow surface modification [67]. A very important feature of the mesoporous silica is that by using different moieties or functional group, the release can be controlled by different external phenomena. Some of the stimuli that can be applied are temperature, redox reactions, enzymes, radiation, pH levels, or reducing agents [68, 69] as shown in figure 19 [66].

The stimuli-responsive release systems are important characteristics of ordered mesoporous materials that make them very competent as drug delivery agents. Thus we see how on demand release and retention of the therapeutic substances is possible by designing mesopore channels in a way that they will respond to externally applied stimuli. Thus we can say that mesoporous channels act as smart carriers, where the drugs can be stored and thereby prevented from deterioration and also released when needed in a controlled manner. In Section 11.6.5, we will discuss about further improvements and future prospects that might be the next step in the development.

11.6.5 Future of Bioceramics

We have already seen how MENs overpower MNs in the two case studies involving ovarian cancer cells and HIV drug; we now propose the idea of bioceramics with MENs. MENs are more superior to MNs in every aspect. It opens up a whole area of research that involves substituting the MNs in the core as mentioned in Ref. [67] with MENs and then studying their characteristics. If MENs were to be used instead of MNs, it would be an efficient alternative to the technology already being used. Further research is required to understand how MENs can be substituted for MNs and to what extent they will improve the efficacy of the existing methodology. However, if achievable, it would mean more number of applications for this promising type of nanomedicine.

11.7 CONCLUSION

We have seen in this chapter how nanoparticles can augment the drug transmission efficacy. There are two types of nanoparticles discussed in this chapter—MNs and MENs. MNs can be broadly described as nanoparticles that can be controlled by magnetic fields. Under the presence of magnetic field, their magnetic moment can be changed. MENs are very similar to MNs, in the sense that they too can be controlled using magnetic fields.

The fundamental difference between MNs and MENs is the presence in the latter of the quantum-mechanically caused ME effect that enables intrinsic coupling between the magnetic spin and the electric dipole. Consequently, energy-efficient and dissipation-free remote control of the intrinsic charge distribution in the MENs (and consequently, control of the bonding force between the MENs and the drug) can be enabled via application of an external magnetic field. Because of the intrinsic ME effect, even if the drug is strongly bonded to the MEN carriers (as required for high-efficacy delivery), it can be fully released at the target location via application of a local magnetic field with a strength above a certain threshold defined by the ME effect. For comparison, the drug release process using the conventional MNs is not controlled at the same fundamental level but instead, is based on an irreversible energy dissipation process caused by an external a.c. magnetic field. In one MN implementation, superpara MNs and drug molecules are incorporated into temperature-sensitive synthetic polymers or other biomaterials that release the drug as the nanoparticles are heated under the field exposure. In other implementations, superpara MNs can be coated with intermediate linkers (tailored to specific load molecules), embedded in a ferro-gel, or directly connected to the load molecules. In either of these conventional cases, the release mechanism with MNs is based on extrinsic phenomena such as, mechanical deformation and/or heat dissipation that affects the release kinetics and are triggered by a relatively high frequency a.c. field (in the range of hundreds of kilohertz or above), consuming a substantial amount of power (in the kilowatt range). On the contrary, the MEN-triggered release process is achieved at the intrinsic level and does not require any intermediate materials. It is dissipation free and extremely energy efficient. The release with MENs can be triggered by an a.c. magnetic field at a relatively low frequency (below 100 Hz) and even at a d.c. field provided the field strength is above a certain threshold value, as described later in more detail, with power consumption in the sub-watt range.

One can argue that MENs may also experience the potential problem of having a relatively large fraction of the nanoparticles (together or without the drug) trapped in the reticuloendothelial system. One solution to this problem is to apply a "zig-zag" shaped time-varying field profile to move the nanoparticles through the system. Then, besides the force along the main delivery path, there is a significant "jolting" force normal to this direction. This "jolting" force ensures that the nanoparticles don't become trapped in the fibers of the reticular connective tissue. However, application of the strongly inertial "zig-zag" force to the conventional MNs can also result in a significant loss of the drug. On the contrary, with MENs, the physical bond between the drug and the nanoparticle can be engineered to be adequately strong to avoid any loss until a command to release is given via an external field.

Learning from this superiority of MENs over MNs, we conducted three experiments using the former. The first experiment used MENs as carriers of the anti-HIV drug AZTTP for its on-demand delivery. The results proved the proposed hypothesis that the drug release can be enhanced to a greater extent when a.c. field is applied as compared to a d.c. field. The experiment also showed how the d.c. magnetic field with a spatial gradient could be used to direct such nanoformulations to the targeted cells, while the drug upon delivery can be effectively released on demand via an

external a.c. magnetic field at 1000 Hz. The second experiment proved our other hypothesis that the MENs loaded with the drug PTX could serve as high-specificity remotely controlled delivery nanosystems to treat EOC through electroporation effects induced by MENs in the vicinity of the cancer cell membranes when exposed to an external magnetic field—an effect at the nanoscale which led to coining of the term "Nano-Electroporation." The third experiment shows the potential applications of MENs ranging from deep brain stimulation for treating neurodegenerative diseases to nanotechnology tools to understand the brain. An *in vivo* study on imprinting control region mice is conducted to show that MENs may be able to address this subject. Due to the ME effect, these "smart" nanoparticles can directly couple intrinsic electric field-driven processes with external magnetic fields for providing another dimension to control the neural activity.

As a final remark, we would like to emphasize that although MENs indeed have a unique advantage of serving in both roles, (i) as regular MNs for drug delivery (via application of a d.c. remote field with a spatial gradient) and (ii) as drug release nanoscale sites, their capability of an on-demand drug release by application of a low remote magnetic field in an extremely low-frequency range is unprecedented and therefore is a merit of its own (even without exploiting the drug delivery feature). In general, this physically controlled release method could be complementary to any other drug delivery and tagging mechanism, whether it is physical or chemical. Although AZTTP is used in the experiment, the ability of on-demand drug release by a MEN nano-carrier discovered in this study is also relevant to the treatment of other diseases such as many CNS diseases, cancer and others where deep-tissue high-efficacy drug delivery at the subcellular level is key. MENs can also have a convincingly bright future in the field of bioceramics. We have already seen the integration of MNs with silica to form mesoporous structures that are highly stable and can act as smart carriers of drugs. They can be used for on-demand release of drugs based on applied stimuli like temperature, pH, and light. By using MENs instead of MNs in the cores of these structures, we might open new possibilities for more efficient system of drug delivery and drug release.

REFERENCES

[1] Fu, B., Lu, R., Gao, K., Yang, Y., Wang, Y. Magnetoelectric coupling in multiferroic $BaTiO_3$–$CoFe_2O_4$ composite nanofibers via electrospinning. EPL 111 (1), 17007 (2015).

[2] Mocroft, A., Lundgren, J. D. Starting highly active antiretroviral therapy: why, when and response to HAART. J. Antimicrob. Chemother. 54 (1), 10–13 (2004).

[3] Wang, X., Chai, H., Lin, P. H., Yao, Q., Chen, C. Roles and mechanisms of HIV protease inhibitor Ritonavir and other anti-HIV drugs in endothelial dysfunction of porcine pulmonary arteries and human pulmonary artery endothelial cells. Am. J. Pathol. 174, 771–781 (2009).

[4] Saxena, S. K., Tiwari, S., Nair, M. P. N. A global perspective on HIV/AIDS. Science 337, 789 (2012).

[5] Thomas, S. A. Anit-HIV drug distribution to the central nervous system. Curr. Pharm. Des. 10, 1313–1324 (2004).

[6] Tong, R. Hemmati, H. D., Langer, R., Kohane, D. S. Photoswitchable nanoparticles or triggered tissue penetration and drug delivery. J. Am. Chem. Soc. 134, 8848–8855 (2012).

[7] Cheong, I., Huang, X., Bettegowda, C., Diaz Jr., L. A., Kinzler, K. W., Zhou, S., Vogelstein, B. A bacterial protein enhances the release and efficacy of liposomal cancer drugs. Science 314, 1308–1311 (2006).

[8] Senyei, A., Widder, K., Czerlinski, C. Magnetic guidance of drug carrying microspheres. J. Appl. Phys. 49, 3578–3583 (1978).

[9] Derfus, A. M., von Maltzahn, G., Harris, T. J., Duza, T., Vecchio, K. S., Ruoslahti, E., Bhatia, S. N. Remotely triggered release from magnetic nanoparticles. Adv. Mater. 19, 3932–3936 (2007).

[10] McBain, S. C., Yiu, H. P., Dobson, J. Magnetic nanoparticles for gene and drug delivery. Int. J. Nanomed. 3(2): 169–180 (2008).

[11] Timko, B. P., Dvir, T., Kohane, D. Remotely triggerable drug delivery systems. Adv. Mater. 22, 4925–4943 (2010).

[12] Guduru, R., Nair, M., Liang, P., Hong, J., Sagar, V., Khizroev, S. Externally-controlled on-demand release of anti-HIV drug AZTTP using magneto-electric nanoparticles as carriers. Nat. Commun. 4, 1707–1713.

[13] Xie, S., Ma, F., Liu, Y., Li, J. Multiferroic CoFe2O4-Pb(Zr0.52Ti0.48)O3 core-shell nanofibers and their magnetoelectric coupling. Nanoscale 3, 3152–3158 (2011).

[14] Eerenstein, W., Mathur, N. D., Scott, J. F. Multiferroic and magnetoelectric materials. Nature 442, 759–765 (2006).

[15] Corral-Flores, V., Bueno-Baqu, D., Ziolo, R. F. Synthesis and characterization of novel CoFe2O4-BaTiO3 multiferroic core-shell-type nanostructures. Acta Mater. 8(3), 764–769 (2010).

[16] Litvinov, J., Nasrullah, A., Sherlock, T., Wang, Y., Ruchhoeft, P., Wilson, R. C. High-throughput top-down fabrication of uniform magnetic particles. PLoS ONE 7(5), e37440 (2012).

[17] Guduru, R., Liang, P., Runowicz, C., Nair, M., Atluri, V., Khizroev, S. Magneto-electric nanoparticles to enable field-controlled high-specificity drug delivery to eradicate ovarian cancer cells. Sci. Rep. 3, 2953–2960.

[18] Zhang, Y., Deng, C., Ma, J., Lin, Y., Nan, C.-W. Enhancement in magnetoelectric response in $CoFe_2O_4$–$BaTiO_3$ heterostructure. Appl. Phys. Lett. 92, 062911–062913 (2008).

[19] Lin, Y., Cai, N., Zhai, J., Liu, G., Nan, C.-W. Giant magnetoelectric effect in multiferroic laminated composites. Phys. Rev. B 72, 012405 (2005).

[20] Mahajan, R. et al. Magnetoelectric effect in cobalt ferrite-barium titanate composites and their electrical properties. Pramana 58, 1115–1124 (2002).

[21] Harker, W. G., Sikic, B. I. Multidrug (pleiotropic) resistance in doxorubicin-selected variants of the human sarcoma cell line MES-SA. Cancer Res. 45, 4091–4096 (1985).

[22] Gosland, M. P., Lum, B. L., Sikic, B. I. Reversal by cefoperazone of resistance to etoposide, doxorubicin, and vinblastine in multidrug resistant human sarcoma cells. Cancer Res. 49, 6901–6905 (1989).

[23] Chen, G. et al. Multidrug-resistant human sarcoma cells with a mutant P-glycoprotein, altered phenotype, and resistance to cyclosporins. J. Biol. Chem. 272, 5974–5982 (1997).

[24] Kesarwani, P., Tekade, R. K., Jain, N. Spectrophotometric estimation of paclitaxel. Int. J. Adv. Pharm. Sci. 2, 29–32 (2011).

[25] Peterson, G., Barnes, S. Genistein inhibition of the growth of human breast cancer cells: independence from estrogen receptors and the multi-drug resistance gene. Biochem. Biophys. Res. Commun. 179, 661–667 (1991).

[26] MacDiarmid, J. A. et al. Sequential treatment of drug-resistant tumors with targeted minicells containing siRNA or a cytotoxic drug. Nat. Biotechnol. 27, 643–651 (2009).

[27] Marblestone, A. H. et al. Physical principles for scalable neural recording. Front. Comput. Neurosci. 7, 137 (2013).

[28] Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. U. S. A. 81, 3088–3092 (1984).

[29] De Kramer J. The electrical circuitry of an olfactory sensillum in *Antheraea polyphemus*. J. Neurosci. 5, 2484–2493 (1985).

[30] Srikanth, M., Kessler, J. A. Nanotechnology-novel therapeutics for CNS disorders. Nat. Rev. Neurol. 8, 307–318 (2012).

[31] Yue, K. et al. Magneto-electric nano-particles for non-invasive brain stimulation. PLoS ONE 7, e44040 (2012).

[32] Brambilla, D. et al. Nanotechnologies for Alzheimer's disease: diagnosis, therapy, and safety issues. Nanomedicine 7, 521–540 (2011).

[33] Kringelbach, M. L., Jenkinson, N., Owen, S. L., Aziz, T. Z.. Translational principles of deep brain stimulation. Nat. Rev. Neurol. 8, 623–635 (2007).

[34] Stefani, A. et al. Bilateral deep brain stimulation of the pedunculopontine and subthalamic nuclei in severe Parkinson's disease. Brain 130, 1596–1607 (2007).

[35] Schlaepfer, T. E. et al. Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. Neuropsychopharmacology 33, 368–377 (2008).

[36] Barker, A. T., Jalinous, R., Freeston, I. L. Non-invasive magnetic stimulation of human motor cortex. Lancet 325, 1106–1107 (1985).

[37] Pascual-Leone, A. Transcranial magnetic stimulation: studying the brain–behaviour relationship by induction of 'virtual lesions'. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 354, 1229–1238 (1999).

[38] Nitsche, M. A., Paulus, W. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. Neurology 57, 1899–1901 (2001).

[39] Boggio, P. S. et al. Effects of transcranial direct current stimulation on working memory in patients with Parkinson's disease. J. Neurol. Sci. 249, 31–38 (2006).

[40] Fregni, F., Pascual-Leone, A. Technology insight: noninvasive brain stimulation in neurology—perspectives on the therapeutic potential of rTMS and tDCS. Nat. Clin. Pract. Neurol. 3, 383–393 (2007).

[41] Yoo, S. S. et al. Focused ultrasound modulates region-specific brain activity. Neuroimage 56, 1267–1275 (2011).

[42] Huang, Y. Z. et al. Consensus: new methodologies for brain stimulation. Brain Stimul. 2, 2–13 (2009).

[43] Coenen, A. M. Neuronal activities underlying the electroencephalogram and evoked potentials of sleeping and waking: implications for information processing. Neurosci. Biobehav. Rev. 19, 447–463 (1995).

[44] Zhao, Z., Zhou, Z., Bao, J. Octapod iron oxide nanoparticles as high-performance T2 contrast agents for magnetic resonance imaging. Nat. Commun. 4, 2266 (2013).

[45] Corral-Flores, V., Bueno-Baques, D., Ziolo, R. Synthesis and characterization of novel $CoFe_2O_4$–$BaTiO_3$ multiferroic core-shell-type nanostructures. Acta Mater. 58, 764–769 (2010).

[46] Kitagawa, Y., Hiraoka, Y., Honda, T., Ishikura, T., Nakamura, H., Kimura, T. Low-field magnetoelectric effect at room temperature. Nat. Mater. 9, 797–802 (2010).

[47] Luo, W. et al. Room-temperature simultaneously enhanced magnetization and electric polarization in $BiFeO_3$ ceramic synthesized by magnetic annealing. Appl. Phys. Lett. 94, 202507 (2009).

[48] Nair, M., Guduru, R., Liang, P., Hong, J., Sagar, V., Khizroev, S. Externally controlled on-demand release of anti-HIV drug using magneto-electric nanoparticles as carriers. Nat. Commun. 4, 1707 (2013).

[49] Guduru, R., Khizroev, S. Magnetic field-controlled release of paclitaxel drug from functionalized magnetoelectric nanoparticles. Part. Part. Syst. Char. 31(5), 605–611 (2013).

[50] Calvo, P. et al. Long-circulating PEGylated polycyanoacrylate nanoparticles as new drug carrier for brain delivery. Pharm. Res. 18(8), 1157–1166 (2001).

[51] Guduru, R., Liang, P., Runowicz, C., Nair, M., Alturi, V., Khizroev, S. Magnetetolectric nanoparticles to enable field-controlled high-specificity drug delivery to eradicate ovarian cancer cells. Sci. Rep. 3, 2953 (2013).

[52] Eskandar, E., Rabinov, J., Barker, F. G. Tracking neural stem cells in patients with brain trauma. N. Engl. J. Med. 355(22), 2376–2378 (2006).

[53] Kim, D. K., Zhang, Y., Kehr, J., Klason, T., Bjelke, B., Muhammed, M. Characterization and MRI study of surfactant-coated superparamagnetic nanoparticles administrated into the rat brain. J. Mag. Mag. Mater. 225, 256–261 (2001).

[54] Goodwill, P. W., Scott, G. C., Stang, P. P., Conolly, S. M. Narrowband magnetic particle imaging. IEEE Trans. Med. Imaging 28, 1231–1237 (2009).

[55] Weizenecker, J., Gleich, B., Rahmer, J., Dahnke, H., Borgert, J. Three-dimensional real-time in vivo magnetic particle imaging. Phys. Med. Biol. 54, L1 (2009).

[56] Yassa, M. A., Muftuler, L. T., Stark, C. E. Ultrahigh-resolution microstructural diffusion tensor imaging reveals perforant path degradation in aged humans in vivo. Proc. Natl. Acad. Sci. U. S. A. 107, 12687–12691 (2010).

[57] Khizroev, S. K., Bain, J. A., Kryder, M. H. Considerations in the design of probe heads for 100 Gbit/in 2 recording density. IEEE Trans. Magn. 33, 2893–2895 (1997).

[58] Tsai, C. P., Chen, C. Y., Hung, Y., Chang, F. H., Mou, C. Y. J. Mater. Chem. 19, 5737 (2009).

[59] Langer, R. Drug delivery and targeting. Nature 392, 5 (1998).

[60] Duncan, R. The drawing era of polymer therapeutics. Nat. Rev. Drug Discov. 2, 347 (2003).

[61] Kawashita, M., Tanaka, M., Kokubo, T., Inoue, Y., Yao, T., Hamada, S., Shinjo, T. Preparation of ferrimagnetic magnetite microspheres for in situ hyperthermic treatment of cancer. Biomaterials 26, 2231 (2005).

[62] Philipse, A. P., Vanbruggen, M. P. B., Pathmamanoharan, C. Magnetic silica dispersions: preparation and stability of surface-modified silica particles with a magnetic core. Langmuir 10, 92 (1994).

[63] Stober, W., Fink, A., Bohn, E. Controlled growth of monodisperse silica spheres in the micron size range. J. Colloid Interface Sci. 26, 62 (1968).

[64] Yoon, T. J. et al. Specific targeting, cell sorting, and bioimaging with smart magnetic silica core-shell nanomaterials. Small 2, 209 (2006).

[65] Hu, S. H., Liu, D. M., Tung, W. L., Liao, C. F., Chen, S. Y. Surfactant-Free, Self-Assembled PVA-Iron Oxide/Silica Core-Shell Nanocarriers for Highly Sensitive, Magnetically Controlled Drug Release and Ultrahigh Cancer Cell Uptake Efficiency. Adv. Funct. Mater. 18, 2946–2955 (2008).

[66] Vallet-Regí, M., Ruiz-Hernández, E. Bioceramics: from bone regeneration to cancer nanomedicine, Adv. Mater. 23, 5177–5218 (2011).

[67] Mizoshita, N., Tani, T., Inagaki, S. Syntheses, properties and applications of periodic mesoporous organosillicas prepared from bridged organosilane precursors. Chem. Soc. Rev. 40, 789 (2011).

[68] Coti, K. K. et al. Mechanised nanoparticles for drug delivery. Nanoscale 1, 16 (2009).

[69] Park, K. Controlled drug delivery systems: past forward and future back. J. Controlled Release 190, 3–8 (2014).

12

DNA COMPUTATION IN MEDICINE

Noam Mamet and Ido Bachelet

Faculty of Life Sciences and The Institute of Nanotechnology and Advanced Materials, Bar Ilan University, Ramat Gan, Israel

12.1 BACKGROUND FOR THE NON-BIOLOGIST

In this chapter, the DNA molecule in the different works is being used not as an hereditary component but as a raw material, as a building block, and as a molecular tool, using its unique traits and the research tools developed in years of study done around this molecule. Some advanced techniques will be discussed later on, but some prior knowledge is needed in order to understand the molecular algorithms and the DNA computation designs described in the rest of the chapter.

DNA is the hereditary component; this molecule encodes the genetic information for living organisms. It is usually found in a double helix structure that consists of two chain-like biopolymers (large molecules composed of many repeated subunits) strands coiled on each other. The basic units of the biopolymer are called "nucleotides" (nt). Four different types of nucleotides can be found in the DNA molecule, either guanine (G), cytosine (C), adenine (A), or thymine (T). The nucleotides are linked together to form a strand of DNA, and their different sequence encodes the genomic data. As noted, the DNA consists of two strands that complement each other in a specific manner according to the Watson–Crick pairing rules: G in one strand will complement C in the other strand and vice versa, and in the same manner A and T will complement each other [1]. These basic units are often referred to as "bases" or when in double-stranded form (dsDNA) "base pairs" (bp). The complementation of the two strands creates a backup system for the data encoded in the DNA, as one strand encodes the data and the other is its template. The encoding strand will often be referred to as the sense strand and the template as the antisense strand (Fig. 12.1).

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

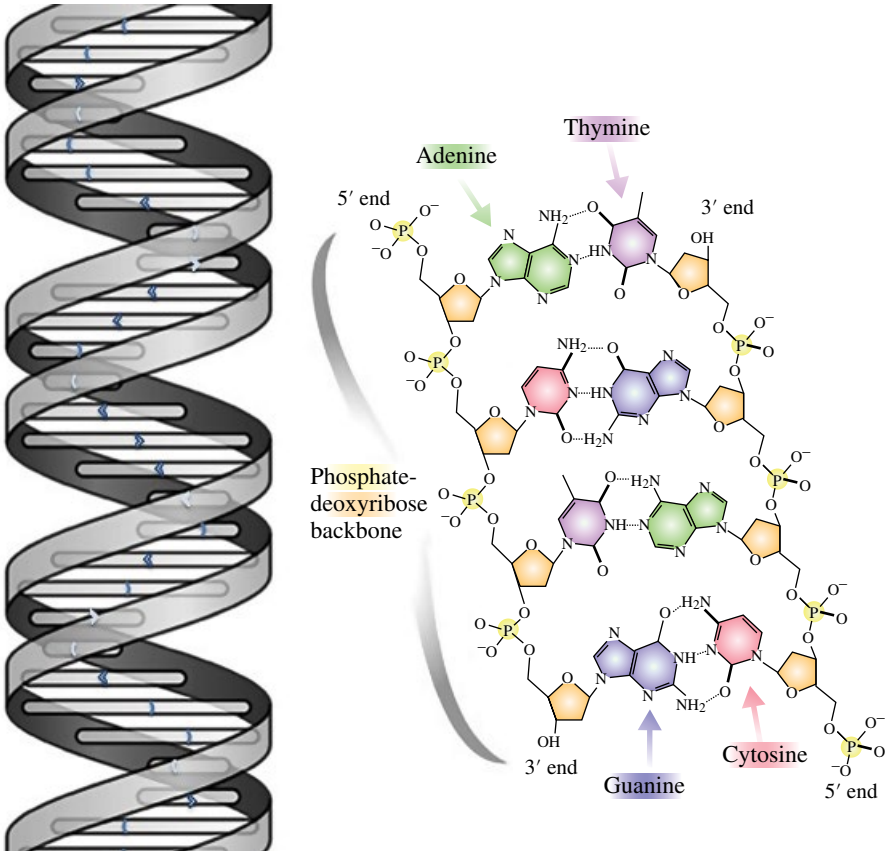


FIGURE 12.1 Left: DNA molecule double helix structure. Right: Watson–Crick pairing rules and the DNA chemical structure [2].

DNA is not always found in its double-stranded (dsDNA) form though it will aspire to be, and will attach to its complementary strand if there is one to be found. For some applications, we intentionally keep the DNA in a single-stranded form (ssDNA) and for some processes the double helix will unzip, going from double strand to two single strands in a reversible process. One process encapsulating on the DNA's "desire" to go back to its double-stranded form is the polymerase chain reaction (PCR) process. The process is done in labs in order to amplify a sample of DNA, that is, to synthesize copies of the DNA strands that are found in the sample. In this process, we use an enzyme called "polymerase," which can synthesize a complementary strand if a strand is found to be used as a template. It can't do it from scratch though, it needs a start to build on. This will be provided by a short complementary DNA strand called a "primer." The primer will hybridize to the template and will be used as a foundation to build on. Then the polymerase can go on and incorporate new nucleotides according to the Watson–Crick pairing rules to complement the template.

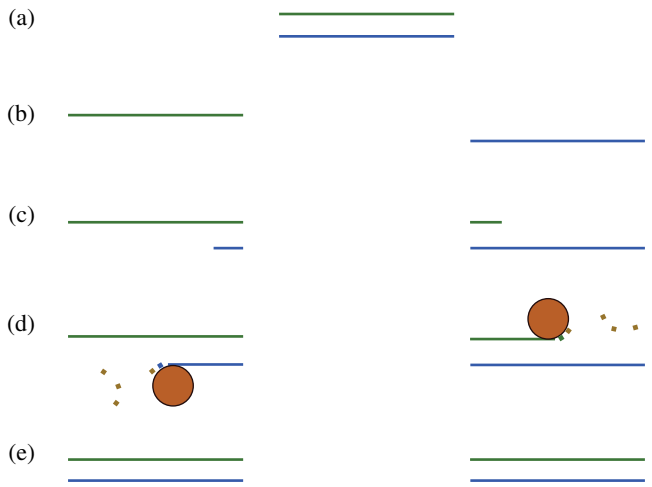


FIGURE 12.2 Schematics of one iteration in a PCR process: (a) dsDNA, (b) denaturation of the dsDNA into two ssDNA, (c) annealing of complementary primers (short lines) to the ssDNA strands, (d) polymerase (circles) elongating the primers by incorporating new nucleotides (little squares) complementary to the template strand, and (e) at the end of every iteration of a PCR process, we double the copies of DNA existing in the sample.

Since the sample DNA will usually be in a dsDNA form, the short primer will not be able to sit on its template. For this reason, the first stage in the PCR process will be to unwind the two complementary strands using high temperature so they could be free and ready for the primer. In the next step, a primer will hybridize to each of the strands, allowing the polymerase to elongate them. At the end of the cycle, we will have two copies of the original dsDNA. This process runs in iterations, doubling the amount of DNA in each round (Fig. 12.2).

Like the polymerase, there are other enzymes that are used to manipulate the DNA molecule. One important family of enzymes is the restriction enzymes. Those will split a DNA molecule in a sequence-specific manner. Different restriction enzymes will look for different sequences; those sequences are referred to as the enzyme recognition site. Some restriction enzymes will break the DNA chain evenly in both strands leaving a blunt end; others will leave one strand longer than the other and overhanging. This overhanging situation is also referred to as sticky ends since the longer ends of the two parts of the broken DNA molecule will still complement each other and will tend to stick together. The opposite function can be achieved using ligase—an enzyme that is used for concatenation of two molecules of DNA.

One common technique to analyze the lengths of strands in a DNA sample is gel electrophoresis. In this method, the sample is loaded into a well found in the gel and then, subjected to an electric field, the DNA strands are pulled through the gel matrix. The DNA molecule has a negative charge, and so it will travel in the gel toward the positive electrode. Longer strands will move slower in the matrix, while shorter strands will move faster making the sample separate according to the different strand lengths in the sample.

12.2 INTRODUCTION

In the past decades, DNA technology had advanced tremendously. The tractability of Watson–Crick pairing provides control over the molecule at the nanoscale. Its predictable features enable us, by nucleotide sequence selection, to program its behavior and to design reproducible complex cascades of events. These characteristics drove the field of DNA technology into a few different directions intersecting here.

12.2.1 Toehold-Mediated Strand Displacement

The toehold-mediated strand displacement is a process in which a strand can take the place of another strand already found in a double-stranded formation [3]. When two strands complement the same sequence, they compete with each other for the same binding site. In case one strand has a longer complementary sequence, it has an advantage. Even if the shorter strand is already hybridized, it will be replaced by it. The exposed single strand left by the shorter sequence is referred to as a toehold. Through the toehold, the process is initiated, as the longer strand will hybridize to it. Then, through branch migration process it will progress, releasing the shorter strand and taking its place (Fig. 12.3). The toehold can be easily blocked by a complementary strand or by making it not accessible, concealed in a hairpin structure, where a strand is folded on itself, and thus making it unreactive. The ability to design and plan the activation and deactivation of the toehold-mediated strand displacement permits a precise control over timing and scale of reactions [4–6].

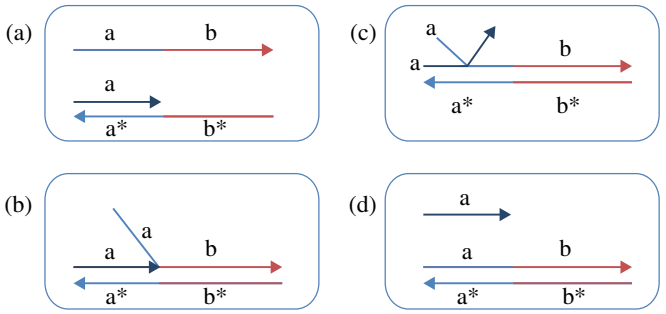


FIGURE 12.3 Toehold-mediated strand displacement. DNA strands are represented in lines with an arrowhead. Domains that correspond with a distinct nucleotide sequence are noted with a letter and their complementary domains are noted with the same letter marked with a small star. (a) Strand ab has two domains marked as "a" and "b." Strand a is attached to its complementary sequence "a*" in strand a*b*, while domain 'b*' is left exposed. (b) Strand ab attaches to strand a*b* through the previously exposed "b*" domain, leaving its "a" domain unattached. This partial attachment is referred to as a toehold. (c) Mediated by the toehold, strand ab pushes strand a, peeling it off and taking its place on the a*b* strand. (d) Eventually, strand a is completely detached, while strand ab is fully hybridized to strand a*b* through their two complementary domains.

12.2.2 Aptamers

RNA or a single-stranded DNA molecule can assume a three-dimensional (3D) conformation that specifically binds to a target. Aptamers can be selected such that they will have a specific affinity to a wide range of targets, whether it's a small molecule, a protein, or a complex target as the surface of a cell. The applications of aptamers are similar to those of antibodies; but once the sequence of an aptamer is known, it is very easy to synthesize and can be amplified in a simple PCR process. Aptamers are being used as molecular tools, as sensors [7], for diagnostics and in therapeutics [8].

For a specific target, there is a certain probability to find an aptamer. That is, if a large enough pool of different strands would be scanned, it is reasonable to assume we can find for almost every target the right sequence and conformation to have affinity to it. One method of finding such molecules for a specific target is the systematic evolution of ligands for exponential enrichment [9] (SELEX). A large pool of randomers are generated first; each randomer has a unique random sequence that is flanked by two known sequences common to all randomers in the pool. Randomers are then incubated with the targets to enable attachment to it. Then a washing step is done to get rid of as much unrelated sequences as possible. Then an amplification step is done yielding an enriched new pool to be incubated with the target for another round of selection and enrichment. The process iterates for several more rounds. At the end of the process, the selected population of strands is much more homogeneous and, hopefully, contains sequences with high affinity to the target (Fig. 12.4).

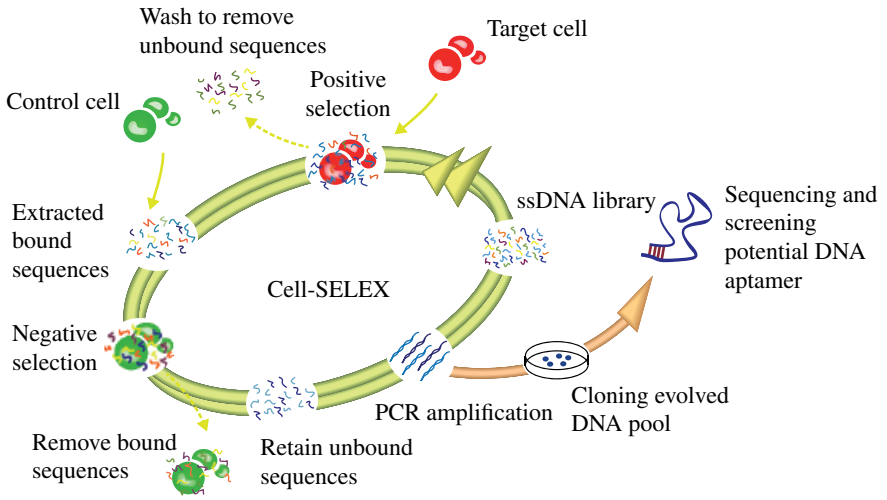


FIGURE 12.4 Cell selex cycle schematics. Starting from a random ssDNA library, going through incubations for first positive and then negative selection. Hopefully, after the selection steps, much of the unwanted strands were washed away. The remaining selected population is then going through an amplification step. The output of the amplification step is then used as the ssDNA library for the next round of selections. After a few rounds, the selected population will be sequenced to reveal the identity of the most common strands of the evolved library. Taken from Sefah *et al.* [9].

12.2.3 DNA Origami

The tractability of Watson–Crick pairing enables the use of DNA as a programmable building block for molecular self-assembly applications. Seeman first proposed the concept of fabricating 2D and 3D lattices from DNA, which could enable the parallel construction of milligrams of nanoscale objects with sub-nanometer resolution and accuracy [10]. This technology was further elaborated by the introduction of DNA origami, first demonstrated by Rothemund [11]. This technique uses a "scaffold" DNA strand. The scaffold is a single strand of DNA, typically several thousands of bases long. It is the raw material for the DNA origami. Directing the folding of the scaffold strand to the desired shape is done by hundreds of short strands called "staples," which hybridize to distinct regions in the scaffold strand and induce crossovers that keep the entire structure solid. The folding reaction is carried out by mixing the scaffold and staple strands and subjecting the mixture to a slow temperature annealing ramp. Scaffolded DNA origami is remarkably robust and reproducible, and it allows the fabrication of an astonishing variety of shapes with arbitrary features and geometries (Fig. 12.5).

12.3 IN VITRO COMPUTING

Adleman first demonstrated molecular computation of solutions to combinatorial problems in 1994 [12]. The computational solution shown by Adleman was a limited instance of the Hamiltonian path problem (HPP). In graph theory, HPP refers to

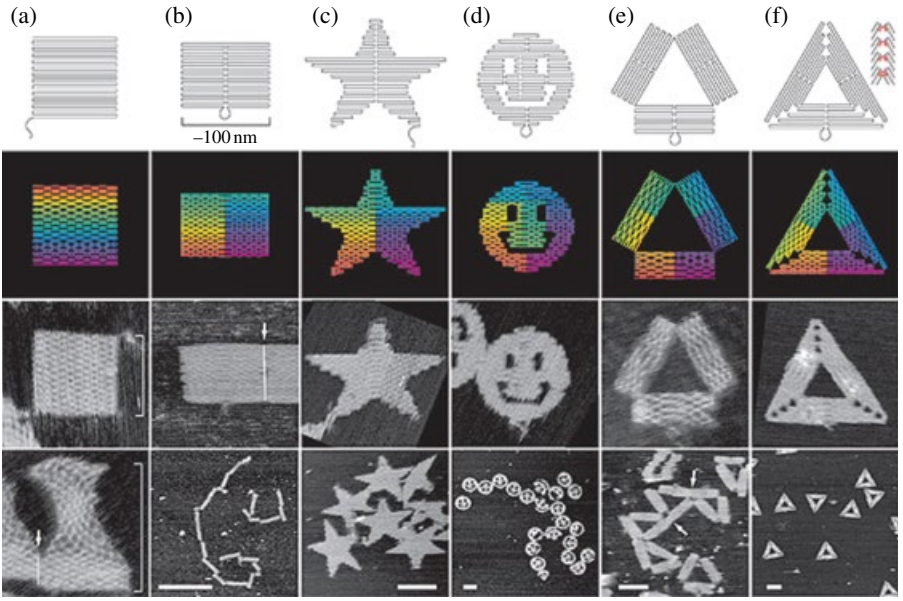


FIGURE 12.5 DNA origami: planed folding path (top row), a graphic model (second row), and AFM images of the DNA origami (two bottom rows). Taken from Rothemund [11].

finding whether there is a path in a given graph visiting each vertex exactly once. These types of problems were proven to belong to a computational complexity theory class of problems called NP-complete problems. There are known algorithms that can solve these types of problems, but as they are NP-complete problems, their complexity in the worst case grows exponentially as the problem grows, making it inefficient and making it inefficient and impractical due to time.

The following steps form nondeterministic algorithms for solving HPPs:

Step 1: Generate random paths through the graph.
Step 2: Keep only those paths that begin with V_{in} and end with V_{out}.
Step 3: If the graph has n vertices, then keep only those paths that enter exactly n vertices.
Step 4: Keep only those paths that enter all of the vertices of the graph at least once.
Step 5: If any paths remain, say "Yes"; otherwise, say "No."

The algorithm was implemented using molecules to solve the HPP for a seven-node graph (Fig. 12.6a). Each of the possible edges in the graph was represented by a 20 nt (nucleotides) long DNA strand. The first 10 nt represented an exit from one vertex and the last 10 nt represented an entrance to another vertex. For each vertex, there was a 20-nt-long DNA strand complementary to the entrance and the exit to the vertex found on the relevant edges (Fig. 12.6b).

All strands were mixed, annealing to each other, randomly generating many different paths welded in a ligation step. By doing so, the first step of the algorithm was carried out. PCR process using primers complementary to the V_{in} and V_{out} representations amplified only those paths that start and end with the predefined starting and

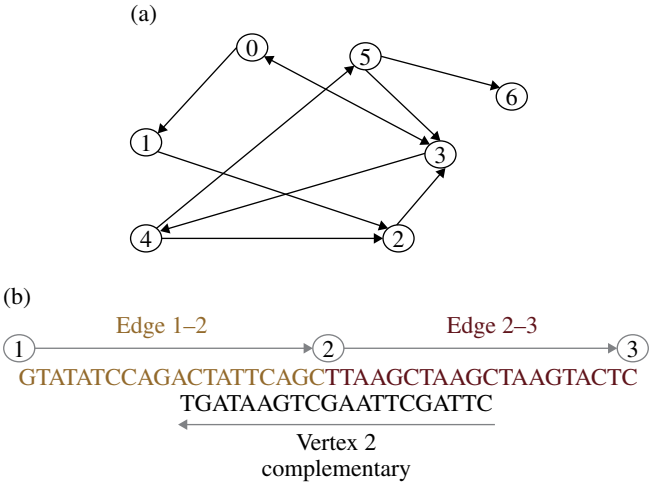


FIGURE 12.6 (a) A scheme of a seven-node-directed graph. Nodes (circles) are numbered according to the Hamiltonian path found for this graph and (b) DNA representation of edges 1–2 and 2–3 annealed to the sequence complementary to the DNA representation of vertex 2.

ending vertices of the graph. The amplified product was then size selected with gel electrophoresis. The 140-bp-selected strands are the ones corresponding to those paths that visit exactly seven vertices on their way as steps 2 and 3 of the algorithm dictates. To carry out step 4, the sense strand of the extracted DNA was checked to hold each of the vertices in a series of purification steps. Magnetic beads were used to keep only those strands that correspond to paths that visit all vertices. In each purification step, complementary strands to one vertex were attached to the beads, holding only strands that contained the specific vertex sequence while the rest washed away. If something is left after step 4, it has to be a Hamiltonian path. A last step of PCR and gel electrophoresis was done in order to get the answer and implement the last step of the algorithm.

Adleman showed the great potential the field stores. It is hard to compare the computation power of his design to the power of a computer; it is not obvious what should count as an operation, each ligation of two edges? Each annealing of one edge to its complementary vertex? Each collusion between two molecules scanning each other to see if they fit?

In order to get some notion of scale, a desktop computer processor speed is in the scale of several gigahertz that is about $(1–4)\times 10^9$ operations per second. In the ligation step of Adelman's implementation, about 4×10^{14} edges were ligated.

Electronic computers work in a linear fashion, and as such they are less efficient for tasks that require scanning of large pools. For such tasks, there can be an advantage in massively parallel processing. Theoretically, the problem and its presented molecular solution can be scaled up without a significant change in the computation time. The change in the amount of vertices will not change the time of the lab operations steps. Nevertheless, up-scaling will present a different problem to this solution. As the problem grows, the exponential growth of possible paths should be translated to an exponential growth of DNA material, which will translate the impractical amount of time into impractical amount of DNA. Other tasks such as multiplying two integers or basically any other algorithm that should be carried out in a linear fashion can be done quite efficiently on electronic computers.

12.3.1 Linear Computing

Turing machine was one of the stimulators for the research that led to modern computers. It is a hypothetical machine that manipulates symbols on a tape according to a set of predefined rules. Though it is simple, using its principles, one can implement any computer algorithm. It is part of a large family of automata that transforms input data of one form to a different form of output data. A finite-state automata is a lesser computational power-wise member of that same family. It accepts a finite string of symbols and runs a computation process accordingly. It has a finite amount of states, an initial one, some transitory states, and some acceptance states, to help define a successful computation. The automaton changes its states in response to a symbol it reads, according to the state it is found in and a set of predefined rules.

Benenson and Shapiro designed such an automaton in which DNA strands and enzymes were used as the building blocks [13]. The input sequences were designed

in the form of dsDNA strands encoding an initiator sequence, a string of symbols, and a terminator state. The processing machinery and the set of rules (hardware and software) were comprised of both a set of sticky end DNA strands, a ligase and FokI restriction enzymes. The FokI cuts the DNA 9nt downstream to its recognition site. At first, the input sequence meets the restriction enzyme; it sits on its recognition site found on the initiator, and trims the double-stranded input, leaving it with exposed sticky ends at the initial state and symbol. Two states and three symbols were designed for this automaton. Each symbol is represented by a sequence of 6 nt. The interaction with FokI leaves four bases of a symbol overhanging, determining the state in which the process is in. The first four bases indicate we are found in state 1 and the last four indicate we are found in state 0. Floating around are software DNA molecules, encoding the rules of transitions. The rules comprised a FokI recognition site, a stretch of DNA in different lengths, and a four-base-long sticky end complementary to the different states and symbols overhanging on the input. The length of the DNA stretch sets the position on the input sequence in which the restriction enzyme will cut next, that is, it determines which four bases will overhang and hence the new state. The processing goes on until a termination symbol is reached. Output detectors complementary to the different states in which the computation had resulted in will ligate to the input creating a distinguished length dsDNA oligo that would be recognized later on in an electrophoresis assay.

This beautiful design implements the idea of linear computation and the flexible abilities of state machine with DNA technology tools. Though theoretically one process could run many different calculations on various inputs in parallel in this case, the content of one tube was running redundantly 10^{13} computation of the same input (Fig. 12.7).

In contrast to the linearity encompassed by the state machine, one computational approach that can benefit from the massively parallel potential of DNA computation is the artificial neural network. It draws inspiration out of biological central nervous system research, using a network of nods to imitate the neural network. The main characteristics of such an approach are adaptive weights and the capability to approximate functions given by the inputs. The adaptive weights indicate the importance of each node in a decision. During the process of learning, the nodes' weights should be tuned . The processing of input is done locally in a parallel fashion.

In 2010, Winfree took a step toward an artificial neural network (ANN) [14]. The ANN was implemented as a strand displacement cascade. One of the main design concepts of this work is a gate element called a "seesaw" (Fig. 12.8); with it, Winfree implements a synapse as it concentrates the functions of input accumulation, threshold, and signal amplification for the next step in line. Different inputs can funnel into the gate, each with a different weight defined by a different number of molecules. First, the input strands encounter a complementary "threshold" strand; only when it is consumed, the input can release the signal waiting annealed to the gate. The releasing is done by strand displacement. The seesaw gate has two specific recognition sites: one for the input signal and the other for the output signal. Those sites are flanking a common sequence to the two. Both input and output strands

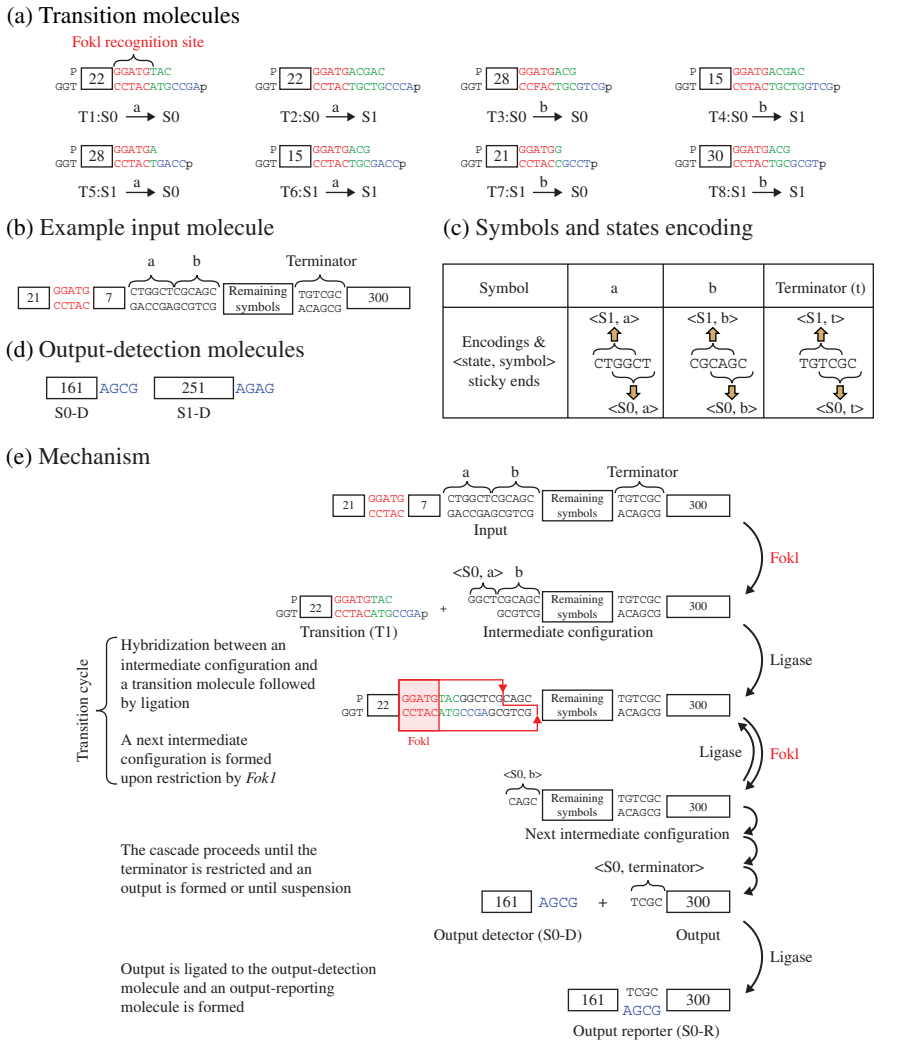


FIGURE 12.7 Benenson's design of a DNA automaton. (a) State transition molecules. Determines the next state according to the current state and symbol. Each transition molecule has a distinct detector in the form of an overhanging sequence to recognize by complementation the right state and symbol on the input, a FokI recognition site (dsDNA sequence, *GGATG-CCTAC*) to recruit the restriction enzyme and a spacer (dsDNA of a changing length that separates between the detector and the FokI recognition site) to determine the FokI cutting location within the input and thus determine the next state. (b) An input molecule basic structure. (c) The sequences encoding on the sense strands for the input symbols a, b, and terminator and the overhanging sequences for each of the symbols that represents the different states. (d) Output detection molecule structure. (e) The steps of an input processing using Benenson's design. Taken from Benenson *et al.* [13].

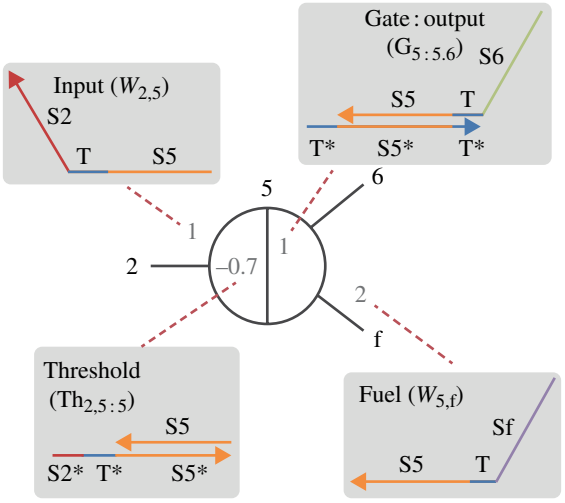


FIGURE 12.8 Abstract diagram of a seesaw gate motif and its DNA implementation. Shaded boxes represent different components in the seesaw gate motif. Lines with arrowheads represent DNA strands. The shading within the lines corresponds to different domains. The unstarred names (e.g., S5) of a domain correspond with the sense strand; the starred names (e.g., S5*) correspond with its complementary, antisense strand. The absolute values of the numbers on the ends of the dotted lines correspond with their component relative concentration at the beginning of the process. s2* complements a small part of S2. Taken from Qian *et al.* [14].

hold a complementary sequence to the recognition site on the gate. Two "wings" flank these complementary recognition elements on the signals. The right "wing" of the input and the left "wing" of the output are similar and complementary to the common sequence on the gate. The seesaw reaction can equally go either way; but as the output detaches, it can keep flowing downstream, becoming the input for the next gate. Another element in the design that helps the downstream flow and amplify signal is in the form of a fuel DNA molecule. The fuel can release an input molecule from the gate after the last one released the output signal. In this way, it prevents the return of the output to the gate and enables the input to act on other gates and to release more output signal molecules. It can be done as the fuel is composed of the same sequence as the output signal except its right wing, and so it can take the place of the output molecule on the gate.

To demonstrate the work of their system, the team used an in silico simulation to train a model of a neural network of their design. The simulation training was used to set the right weights for the neural network so it would be able to recognize a known scientist according to a list of characteristics. The characteristics fed into the network as inputs cascaded downstream and reported the right scientist by displacing a quencher from the scientist representation dye.

12.4 COMPUTATION IN VIVO

12.4.1 Paradigm Shift in Drug Control and Design

Controlling the location and timing of the activity of molecules is a central requirement in nanotechnology. Every cellular process is the result of molecular interactions that are exquisitely regulated in time and space. In contrast, our ability to mimic this level of precise control at the molecular scale is exceedingly limited. This is particularly significant for the field of pharmacology and drug design. The ideal drug acts only on its designated target at the proper timing. Unfortunately, most drugs currently in use have an enormous range of adverse effects due to lack of proper spatiotemporal control [15–18]. In fact, our current therapeutic arsenal includes very effective molecules; however, the lack of ability to properly control them hinders their safe, effective use.

Current improvements in drug special control are mostly achieved through conjugation of the therapeutic molecule to target-specific carriers such as antibodies or interleukins [19], and local implantation of a therapeutic molecule. Temporal control is mostly achieved by embedding the drug in a matrix that hinders and prolongs its dissolution and diffusion, thereby causing the sustained, more stabile distribution of the drug [20].

While definitely representing improvements over conventional, unmodified drug administration, many desired features are not enabled by these techniques. Examples for such features include reversing the availability of a molecule at will, coordinating the activity of two or more, or activating a molecule by a complex combination of biological conditions. Applying this higher level of control over drug administration and design would lead to significant improvements in the safety and efficacy of many drugs in use today. Theoretically, a computer capable of reading biomolecular inputs, and linking its output to drug release, could arguably solve the challenge of drug control.

Thus, maybe the most important advantage of DNA computation is exactly that. A direct interface with biomolecules. This ability to interact directly with molecular signals found in a natural surrounding allows DNA computation devices to read a secreted cell signal or hormone as inputs, scan the surface of cells, or even react to processes inside the cell .

Once we can feel our biology surrounding, it can be analyzed locally to generate a diagnosis or even immediately calculate the best course of action and intervene, for example, by releasing a drug.

A few works have been done to present possible mechanisms that will get inputs from the surrounding. One example to that is the next step in Shapiro's work with his automata [21]. Using its flexibility, it can be implemented to calculate many different things as the state transition rules are being activated according to inputs supplied by the surrounding. An example that was used in the paper showed two states: machine one state indicates Yes—the input we search is found. The other indicates No—the input is absent. The state machine initialized as Yes, and was expecting to get a transition rule if the input was found in the surrounding, will get a transition rule that will keep it in Yes state and enables the release of a quencher and exposing a fluorescent dye.

A different work, by Maria Rudchenko, was set to differentiate between two subpopulations of lymphocytes [22]. The work leans on unique combination of cell surface markers, as usually there is no unique marker specific to only one cell type. The design makes use of antibodies specific to different cell surface markers, covalently connected to different strands of DNA. With the principle of strand displacement, one strand can release another, cascading downstream through all relevant markers. If all markers are present in close proximity, the last strand displacement will report by releasing a quencher. The design is of an AND gate, if one of the markers will be missing, the chain of displacements will stop. In the same way, it can make sure one of the markers, or more, is missing stopping the cascade if the marker is present.

A very similar work was published by Mingxu You who designed a DNA junction referred to as a nano-claw [23]. At the end of one arm attached is the catalytic/reporter molecule annealed to a blocking strand, while the rest of the arms holds aptamers specific to cell surface markers and annealed to their complementary strands. Upon an encounter with its fitting cell surface marker, the aptamer will attach to it, releasing its complementary strand. The released strand is then used to strand displace part of the sequence that blocks the catalytic/reporter arm. In his work, he had successfully treated cancer cells in a specific manner using the nano-claw to direct a photodynamic therapy.

In 2012, Douglas and Bachelet [24] designed the next step in the form of a DNA origami clam-like nanorobot. This DNA nanorobot has the ability to sense its environment and perform computation in a similar fashion to the designs described before with the advantage of carrying a compartmentalized cargo to be exposed at the right location in response to the detection of a target. The DNA nanorobot sensing mechanism uses aptamers specific for the designated target. The aptamer is connected to one side of the clam lip; and while targets are not found, it is annealed to a partially complementary strand of DNA connected to the other lip of the clam-like robot. Thus, it keeps the nanorobot closed and the cargo concealed. In their work, Douglas and Bachelet demonstrated the nanorobot ability to distinguish six different cell lines using different combinations of three well-characterized aptamers as locks. Two locks on each type of robot were used to target a specific combination of keys. On their cargo, a fluorescently dyed antibody against the human HLA-A/B/C, expressed by all six cell lines, was loaded. When a cell line expression pattern encoded the right set of keys, the nanorobot opened and presented its cargo to allow a connection between the labeled antibody and its HLA-A/B/C antigen. To further demonstrate the robot ability to specifically target one cell type in a complex environment, a natural killer leukemia (NKL) cell line was mixed with healthy whole blood leukocytes. Nanorobots loaded with fluorescently labeled anti-CD33 and locked with anti-platelet-derived growth factor (PDGF) aptamers selectively labeled the NKL cells in the mixture with high precision (Fig. 12.9).

Later on Amir and Bachelet [25] took this work one step further as they not only showed computation can be done in a living animal but also presented the possibility to create any logic gate using layers of different designs of robots, thus transforming the nanorobot to a basic computation element.

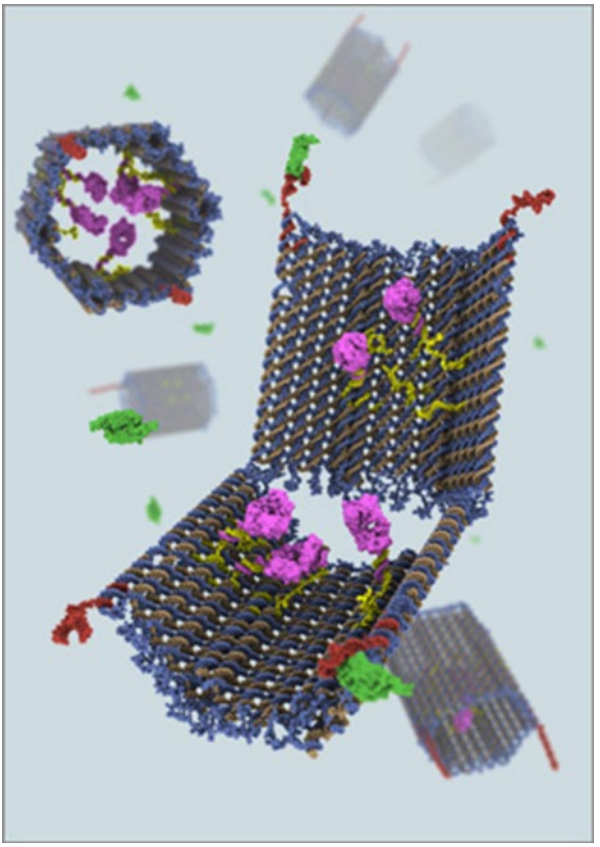


FIGURE 12.9 An illustration of a DNA origami nanorobot.

Following simple generic rules of construction, one can design layers of nanorobots to perform any algorithm. The algorithm can be carried out in a linear fashion if needed or by taking the advantage of molecular computing in a massively parallel process.

One more notable work by Zhen Xie demonstrates an in cell incorporation of computation circuits to evaluate complex cell conditions [26]. Changes in endogenous markers occur in response to different conditions or states. Using the changing patterns, it is possible to classify cells to their states accurately. Harnessing the power of synthetic biology Zhen Xie designed a cell classifier. As a proof of concept, it was made to look for changes in expression patterns of six endogenous miRNA that combines to uniquely identify HeLa cervical cancer cells. An apoptosis process was triggered if a cell was classified to be an HeLa cell. This line of research, using computation elements from within cells, can be used together with the extracellular elements, thereby spreading the reach of the technology farther and the scope of possibilities wider.

12.5 CHALLENGES

Molecular computers can be a powerful tool for diagnosis and therapeutics and can have a profound impact on the field of medicine enabling novel capabilities that cannot be achieved using current technologies. However, there are challenges yet to be solved before molecular computers—possibly in the form of nanorobotics swarms, artificial neural networks, or automatas—become widespread.

In the following years, research will have to look for strategies for the up-scaling of computation abilities while keeping the systems stable and robust. Another step that is yet to be taken is the testing of the systems in mammalian animal models. The complex environment may pose many obstacles to overcome some concern in the immunogenicity of the systems. Toll-like receptor 9 (TLR9), an intracellular mechanism for the detection of CpG non-methylated DNA [27], can exert a potent immune response as it can recognize the molecular computer DNA as a pathogenic agent. DNA is inherently immunogenic, and can cause the eruption of autoimmune diseases such as systemic lupus erythematosus (SLE) [28, 29], when a DNA spillage from damaged cells is not cleared efficiently. The stability of the system can be hindered as well, as nucleases found in the bodily fluids [30] might degrade the system. Some directions to deal with these problems can be by changing the sequence of DNA used. In the case of DNA origami, the scaffold used should be of mammalian origins, as it seems that the immunogenicity of the DNA is diverse and dependent upon sequence and backbone structure. While bacterial DNA is immunologically potent, it seems that mammalian DNA can suppress certain immune responses [31]. To deal with DNAse degradation, incorporation of synthetic nucleotides such as LNA or the modification of oligonucleotides to form a phosphorothioate linkage can be tested.

12.6 GLIMPSE INTO THE FUTURE

Different works are starting to create the critical mass that is the sign of the birth of a field on its own—biocomputing. We are now found in a time that resembles the mid-1940s of electronic computers. The building blocks of the technology are already there and the limits are the boundaries of one's imagination. It is plausible that in the near future drugs will be delivered to their designated target using technologies such as those depicted earlier. Not so far after that temporal control of drugs will be added, and then they will be released according to the body demands and needs. One can imagine the impact such a system would have, for instance, on diabetics, when the insulin will be waiting in their bloodstream to be released when it is needed. Maybe later in time, one generic pill would contain a system that will calculate a diagnosis for various conditions and will intervene in the proper way autonomously, or perhaps when one drug will stop being effective against a mutated cancerous tumor, the system will sense that, and much like our GPS navigators, it will calculate a new route of action and will release a different drug. One day, maybe not in the near future, but then again, maybe not so far away, biocomputing will be everywhere and fully integrated into our lives.

REFERENCES

[1] Watson, J. D. and Crick, F. H. C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171, 737–738 (1953).

[2] DNA. Wikipedia, the free encyclopedia (2015). Available at: https://en.wikipedia.org/w/index.php?title=DNA&oldid=686172400 (accessed on January 13, 2016). "DNA chemical structure" by Madprime (talk contribs)—Own work. The source code of this SVG is valid. This vector image was created with Inkscape. Licensed under CC BY-SA 3.0 via Commons. Available at: https://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg#/media/File:DNA_chemical_structure.svg (accessed on January 13, 2016).

[3] Zhang, D. Y. and Seelig, G. Dynamic DNA nanotechnology using strand-displacement reactions. Nat. Chem. 3, 103–113 (2011).

[4] Zhang, D. Y. and Winfree, E. Control of DNA strand displacement kinetics using toehold exchange. J. Am. Chem. Soc. 131, 17303–17314 (2009).

[5] Yurke, B. and Mills, A. P. Jr. Using DNA to power nanostructures. Genet. Program. Evolvable Mach. 4, 111–122 (2003).

[6] Li, Q., Luan, G., Guo, Q., and Liang, J. A new class of homogeneous nucleic acid probes based on specific displacement hybridization. Nucleic Acids Res. 30, e5 (2002).

[7] Stojanovic, M. N., de Prada, P., and Landry, D. W. Aptamer-based folding fluorescent sensor for cocaine. J. Am. Chem. Soc. 123, 4928–4931 (2001).

[8] Xiang, D. *et al.* Nucleic acid aptamer-guided cancer therapeutics and diagnostics: The next generation of cancer medicine. Theranostics 5, 23–42 (2015).

[9] Sefah, K., Shangguan, D., Xiong, X., O'Donoghue, M. B., and Tan, W. Development of DNA aptamers using cell-SELEX. Nat. Protoc. 5, 1169–1185 (2010).

[10] Chen, J. H. and Seeman, N. C. Synthesis from DNA of a molecule with the connectivity of a cube. Nature 350, 631–633 (1991).

[11] Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. Nature 440, 297–302 (2006).

[12] Adleman, L. M. Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994).

[13] Benenson, Y. *et al.* Programmable and autonomous computing machine made of biomolecules. Nature 414, 430–434 (2001).

[14] Qian, L., Winfree, E., and Bruck, J. Neural network computation with DNA strand displacement cascades. Nature 475, 368–372 (2011).

[15] Cooper, W. O. *et al.* ADHD drugs and serious cardiovascular events in children and young adults. N. Engl. J. Med. 365, 1896–1904 (2011).

[16] Grant, L. M. and Rockey, D. C. Drug-induced liver injury. Curr. Opin. Gastroenterol. 28, 198–202 (2012).

[17] Jukema, J. W., Cannon, C. P., de Craen, A. J. M., Westendorp, R. G. J., and Trompet, S. The controversies of statin therapy: Weighing the evidence. J. Am. Coll. Cardiol. 60, 875–881 (2012).

[18] Ray, W. A., Murray, K. T., Hall, K., Arbogast, P. G., and Stein, C. M. Azithromycin and the risk of cardiovascular death. N. Engl. J. Med. 366, 1881–1890 (2012).

[19] Govindan, S. V. and Goldenberg, D. M. Designing immunoconjugates for cancer therapy. Expert Opin. Biol. Ther. 12, 873–890 (2012).

[20] Moses, M. A., Brem, H., and Langer, R. Advancing the field of drug delivery: Taking aim at cancer. Cancer Cell 4, 337–341 (2003).

[21] Gil, B., Kahan-Hanum, M., Skirtenko, N., Adar, R., and Shapiro, E. Detection of multiple disease indicators by an autonomous biomolecular computer. Nano Lett. 11, 2989–2996 (2011).

[22] Rudchenko, M. *et al.* Autonomous molecular cascades for evaluation of cell surfaces. Nat. Nanotechnol. 8, 580–586 (2013).

[23] You, M. *et al.* DNA 'nano-claw': Logic-based autonomous cancer targeting and therapy. J. Am. Chem. Soc. 136, 1256–1259 (2014).

[24] Douglas, S. M., Bachelet, I., and Church, G. M. A logic-gated nanorobot for targeted transport of molecular payloads. Science 335, 831–834 (2012).

[25] Amir, Y. *et al.* Universal computing by DNA origami robots in a living animal. Nat. Nanotechnol. 9, 353–357 (2014).

[26] Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R., and Benenson, Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. Science 333, 1307–1311 (2011).

[27] Hemmi, H. *et al.* A toll-like receptor recognizes bacterial DNA. Nature 408, 740–745 (2000).

[28] Tan, E. M. Autoantibodies, autoimmune disease, and the birth of immune diagnostics. J. Clin. Invest. 122, 3835–3836 (2012).

[29] Christensen, S. R. *et al.* Toll-like receptor 9 controls anti-DNA autoantibody production in murine lupus. J. Exp. Med. 202, 321–331 (2005).

[30] Nadano, D., Yasuda, T., and Kishi, K. Measurement of deoxyribonuclease I activity in human tissues and body fluids by a single radial enzyme-diffusion method. Clin. Chem. 39, 448–452 (1993).

[31] Pisetsky, D. S. Immune response to DNA in systemic lupus erythematosus. Isr. Med. Assoc. J. 3, 850–853 (2001).

13

GRAPHENE-BASED NANOSYSTEMS FOR THE DETECTION OF PROTEINIC BIOMARKERS OF DISEASE: IMPLICATION IN TRANSLATIONAL MEDICINE

FARID MENAA[1], SANDEEP KUMAR VASHIST[2], ADNANE ABDELGHANI[3], AND BOUZID MENAA[4]

[1] *Department of Nanomedicine, Oncology and Stem Cells, Fluorotronics, Inc., San Diego, CA, USA*

[2] *Hahn-Schickard-Gesellschaft für Angewandte Forschung E.V. (HSG-IMIT), Freiburg, Germany*

[3] *Nanotechnology Laboratory, National Institute of Applied Science and Technology, Carthage University, Charguia Cedex, Tunisia*

[4] *Department of Infection Control, Hymetec SA, Isnes, Belgium*

13.1 INTRODUCTION

13.1.1 History of Graphene

Graphene (G) (Fig. 13.1), a two-dimensional (2D) nanomaterial, is an allotrope (i.e., another form of a chemical element in a same physical state) of carbon that was first described by Boehm in 1962 [1–3]. G represents free-standing 2D honeycomb crystals or single-atom-thick crystallites [4–6]. G can be wrapped up into 0D fullerenes (i.e., fullerene is a molecule of carbon in the form of a hollow sphere, ellipsoid, tube, and many other shapes) [7], rolled into 1D carbon nanotubes (CNTs) [7], or stacked into 3D graphite (GP) [5]. In 2004, Geim and Novoselov used a "Scotch Tape"

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

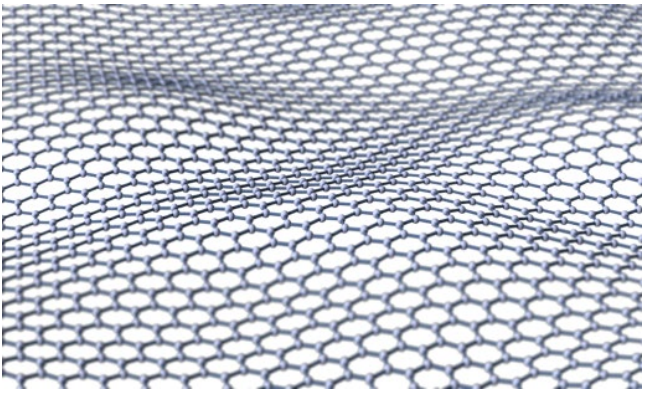


FIGURE 13.1 Two-dimensional structure of graphene.

technique to isolate G from bulk GP, demonstrating that GP is a stack of multi-G sheets [8]. The "scotch tape" method is a micromechanical exfoliation using an adhesive tape to repeatedly split GP crystals into increasingly thinner pieces (i.e., 0.01 thousandth of an inch) in the presence of silicon dioxide (SiO_2), which could be used as a "back gate" electrode to vary the charge density in the extracted nearly neutral (zero-gap) G semiconductor [8, 9]. Since 2004, the physicochemical features (e.g., electronical [6–8, 10–15], optical [16–22], mechanical [23], and thermal [24–28]) as well as the structural properties [29, 30] of G have been extensively studied. Further, since 2008, it is possible to scale up the production of sustainable, efficient, and cost-effective G sheets of quality [5, 31–33]. This can be explained by the development of innovative processes such as exfoliation by dry deposition (i.e., drawing method [8], layer-by-layer (LBL) self-assembly [34, 35], electrochemical reduction of exfoliated G oxide (GO) [36, 37], and epitaxial growth in GP) or in metals as substrates [38–41], particularly via chemical vapor deposition (CVD) [41]. In 2010, Geim and Novoselov were jointly awarded the Nobel Prize in Physics for their groundbreaking experiments regarding G [42].

13.1.2 Utility and Awareness of Graphene and Derivatives in Diagnostic Medicine

Nowadays, free-standing G sheets and derived functionalized carbon nanostructures (e.g., G and GO hybrid nanocomposites) attract many researchers and industrials interested in innovative development of eco-friendly and reliable, in terms of sensitivity, stability, specificity, selectivity, and rapidity, nanoscale-based biological and biochemical sensors (e.g., field-effect transistors (FETs), fluorescence resonance energy transfer (FRET), or chemiluminescence resonance energy transfer (CRET)), because G and G derivatives possess extraordinary intrinsic tunable properties (e.g., electronical, optical, and mechanical) [43–51]. Such powerful nanobioimaging and biomolecular sensors are of high importance in medicine (e.g., earlier, faster, and more reliable detection of diseases such as cancers, fetal

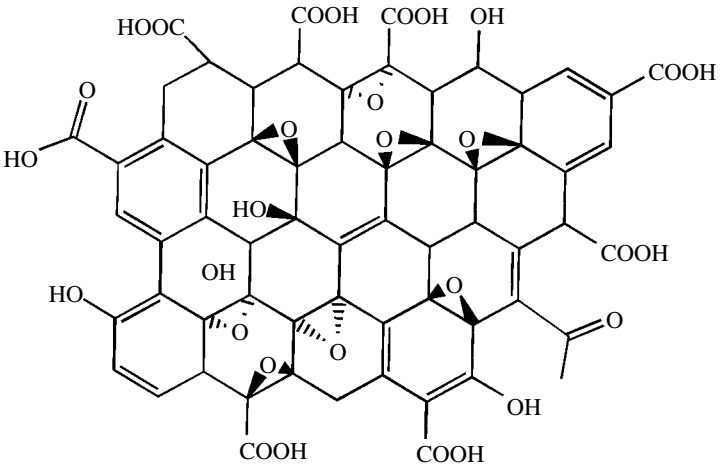


FIGURE 13.2 Two-dimensional structure of graphene oxide.

malformations, hemoglobinopathies), and represented until recently a challenge. Nevertheless, GO (Fig. 13.2), one of the major G derivatives, commonly obtained by oxidation from G using Hummer's method (i.e., combination of $KMnO_4$, sodium nitrate ($NaNO_3$), and H_2SO_4) [52], can be characterized with different electrical conductivity based on the degree of oxidation and its possible chemical surface functionalization.

Moreover, although G and derivatives offer exceptional physicochemical properties and are relatively safe *in vitro*, especially due to constant improvement of surface designs (i.e., physical or chemical surface functionalizations), one should also be aware of their potential cell and systemic nanotoxicity *in vivo* (e.g., G functionalization with heavy metals might cause deleterious effects) [53–59].

This chapter reports the recent advances made in the field of G and major G derivatives-based nanosystems intended for the detection of proteinic biomarker(s) of disease(s). It also provides some perspectives and challenges for more efficient and safer translational medicine (i.e., a bench-to-bedside biomedical and public health research approach that aims to improve the health of individuals by "translating" findings into diagnostic tools, medicines, procedures, policies, and education).

13.2 STRUCTURAL AND PHYSICOCHEMICAL PROPERTIES OF GRAPHENE AND MAIN DERIVATIVES

13.2.1 Structure and Molecular Stability

Indeed, G is a fascinating allotrope of pure sp^2-bonded carbon atoms densely arranged in a regular hexagonal crystal lattice (i.e., an infinite 1, 2, or 3D regular arrangement of points, each of which has identical surroundings) pattern similar to GP but in a single planar sheet [4–6]. G also represents a basic building block for graphitic

materials of all other dimensionalities. It can be wrapped up into 0D fullerenes, rolled into 1D nanotubes, or stacked into 3D GP [5, 7]. However, G sheets in these devices have irregular shapes and variable sizes and contain various impurities and defects (i.e., ripples), which are undesirable for applications, questioning the two-dimensionality of G [15]. Interestingly, G can self-repair holes in G sheets when exposed to molecules containing carbon, like hydrocarbons, or when bombarded with pure carbon atoms [59].

The atomic structure of an isolated, single-layer G (SLG) was studied by transmission electron microscopy (TEM) (i.e., a microscopy technique in which a beam of electrons is transmitted through an ultrathin specimen, interacting with the specimen as it passes through). Alternate modes of use allow for the TEM to observe modulations in chemical identity, crystal orientation, electronic structure, and sample-induced electron phase shift as well as the regular absorption-based imaging on G sheets suspended between bars of a metallic grid [60]. G is strong, light (about $0.77\,mg\,m^{-2}$), and flexible due to its molecularly gate-tunable planar structure and harbors a large surface area [29, 61]. The length of its carbon–carbon (C—C) bonds is about 0.142 nm [29]. G sheets stack to form GP with an interplanar spacing of 0.335 nm [62].

Recently, the atomic structure of G was also revealed by atomic force microscopy (AFM) (i.e., a very high-resolution type of scanning probe microscopy (SPM), with demonstrated resolution on the order of fractions of a nanometer, more than a thousand times better than the optical diffraction limit). It is possible to measure with an AFM the roughness of a sample surface at a high resolution, to distinguish a sample based on its mechanical properties such as hardness and roughness and to perform a microfabrication such as atomic manipulation when exploring the interaction of polyelectrolytes such as single-stranded deoxyribonucleotide (ssDNA) with G as a substrate [30]. Interestingly, it was observed that (i) in absence of screening ions deposition, ssDNA specifically interacts with G and not with the SiO_2 substrate, confirming that the binding energy was mainly due to pi (π–π) stacking/electrostatic interaction and that (ii) this later interaction correlates between the amount of deposited DNA and the G layer thickness. For information, the π effect or interaction is a type of noncovalent interaction, the electron-rich π system being able to interact with a metal (cationic or neutral), an anion, another molecule, and even another π system.

Ab initio (i.e., from the beginning) calculations showed that a G sheet is thermodynamically unstable (i.e., tendency to scroll and buckle [63]) on a nanometer scale (size <20 nm or <24,000 carbon atoms) [63], certainly because of its lower-energy state [64].

All together, the overall published data (e.g., in terms of structure and properties) showed that G could represent an ideal molecular substrate.

13.2.2 Electronic Properties

G has received widespread attention due to its extraordinary electronic properties.

Wallace was the first to show that G is a semimetal or zero-gap semiconductor because the dispersion (E–k) relation is linear (or "conical") for low energies near

the six corners of the 2D hexagonal Brillouin zone (i.e., Dirac points), leasing to zero effective mass for electrons and holes (i.e., Dirac relativistic-like spin-1/2 particles aka Dirac fermions or graphinos) [10, 11, 65–67]. The fact that the band gap of G is zero limits applications (e.g., building of logic circuits due to small on/off ratios of G-FETs). However, a recent study [68] reported a direct measurement of the quantum capacitance of G, as a function of gate potential using a three-electrode electrochemical configuration. It showed that (i) the quantum capacitance has a nonzero minimum at the Dirac point and a linear increase on both sides of the minimum with relatively small slopes and that (ii) the charged impurities and ionic concentrations in aqueous solutions would influence the quantum capacitance. This last observation is in agreement—at least partially—with other experimental results from transport measurements, which showed that G has a remarkably high electron mobility at room temperature (i.e., $15{,}000\,cm^2V^{-1}s^{-1}$ and up to $40{,}000\,cm^2V^{-1}s^{-1}$ for G on SiO_2 substrates) [5, 11–15], notably because of (i) low resistivity of the G sheet (i.e., $10^{-6}\,\Omega\,cm^{-1}$) and also because (ii) certain level concentration of ionized impurities (e.g., lowest levels of potassium (K)) and/or chemical dopants (e.g., gaseous species) possibly present on SiO_2 substrates. Thereby, a recent interesting study [69] has reported transport properties of G transistors (G-FETs) in different solvents, showing that the carrier mobility can vary from 2 to 3 orders of magnitude and reach approximately $70{,}000\,cm^2V^{-1}s^{-1}$, depending on the dielectric constant.

Besides, bilayer G (BLG) was shown to exert a quantum hall effect, a tunable band gap [70], and a potential for excitonic condensation [71], making them promising candidates for optoelectronic and nanoelectronic applications (e.g., sensors). One way to synthesize BLG could be via CVD and can produce large-area bilayer regions that almost exclusively conformed to a Bernal stack geometry (i.e., G layers arranged in ABAB sequence as seen by Raman spectroscopy (RS); http://ucrtoday.ucr.edu/1229/2733_0hi) [72].

13.2.3 Optical Properties

As a consequence of the unusual low-energy electronic structure of SLG, the unique optical properties of G are produced by an unexpectedly high opacity for an atomic monolayer in vacuum (i.e., absorbance of 2.3% of white light) [16, 73].

Recently, a G-based Bragg grating (1D photonic crystal) has been fabricated and demonstrated its competence for excitation of surface electromagnetic waves in the periodic structure using prism coupling technique [74]. Besides, G/GO system exhibits electrochromic behavior, allowing electrical tuning of both linear and ultrafast optical properties [17]. Moreover, the optical response of G nanoribbons (i.e., G obtained from cut CNTs) can be tunable into the terahertz (THz) regime by an applied magnetic field [18].

Eventually, due to its nonlinear optical behavior (i.e., G-based saturable absorption under strong excitation over the visible to near-infrared (NIR) region), G can be then widely applied in ultrafast photonics (e.g., sensing applications, mode locking of fiber lasers, microwave, and THz photonic devices) [19–22].

13.2.4 Mechanical and Thermal Properties

G has a breaking strength over a hundred times greater than a hypothetical steel film of the same thickness and a stiffness of 1 terapascal (TPa i.e., 150,000,000 psi) [23].

The interaction of G with an electromagnetic field is surprisingly strong for a one-atom-thick material, and so the Casimir effect (i.e., interaction between any disjoint neutral bodies provoked by the fluctuations of the electrodynamical vacuum in G systems) is of growing research interest [75, 76]. The related van der Waals force (i.e., dispersion force) is also unusual, obeying an inverse cubic asymptotic power law in contrast to the usual inverse quartic [77].

The near-room temperature ballistic thermal conductivity of G is phonon denominated (i.e., a quantum of vibrational energy in a crystalline solid) and isotropic and ranges from $4.84 \pm 0.44 \times 10^3$ to $5.30 \pm 0.48 \times 10^3\,\mathrm{W\,m^{-1}\,K^{-1}}$ [24–28].

13.3 GRAPHENE AND DERIVATIVES-BASED BIOSENSING NANOSYSTEMS AND APPLICATIONS

13.3.1 Current Status

G's unique properties (e.g., electrical and optical) have made it a popular candidate for nanomaterial-based electrochemical and optical biosensing due to its remarkable structural and physicochemical characteristics [78].

While recent biosensors have incorporated CNTs as sensing elements [79–81], a growing body of work has begun to do alternatively the same with the emergent unrolled nanomaterial G in its original or functionalized forms [81–85].

The ongoing considerable interest in functionalizing G films (e.g., development of G and GO hydride nanocomposites by chemical (e.g., decoration with noble metals), physical (e.g., ionic liquid), or biological (doping with certain peptides or proteins)) procedures aims to better control their electronic properties, enhance their binding to other molecules for sensing, and strengthen their interfaces with matrices in a composite material [78, 86–89].

Indeed, the large surface area and excellent electrical conductivity of the extracted G and derivatives (e.g., GO and reduced GO/rGO) allow them to act as an "electron wire" between the redox centers of an enzyme (e.g., horseradish peroxidase (HRP) and glucose oxidase (GOD)) or a protein and an electrode's surface (e.g., glassy carbon electrode (GCE)), thereby facilitating accurate, relatively stable, fast, highly sensitive and selective/specific detection of single or multiplexed biomolecules [90, 91], including peptides or proteins [92–95], protein markers of diseases [48, 84, 88, 96], or even glucose [81].

Interestingly, it has been shown that GO flakes in polymers displayed enhanced photoconducting properties compared to unmodified GO [97]. GO can also act as an ultrahighly efficient quencher (e.g., fluorescence or chemiluminescence quenching), electron transfer facilitator, and can be reduced and/or functionally hybridized with metals, chemicals, and/or biomolecules [90, 98].

The characterization of G, GO, and corresponding functionalized derivatives (i.e., hybrid nanocomposites) as well as their molecular interactions/binding/adsorption can be achieved by several physical techniques. These include the use of AFM, TEM, scanning electron microscopy (SEM), thermogravimetric analysis (TGA), UV-Vis, photoluminescence spectrometry, energy dispersive X-ray spectroscopy (EDS), X-ray photoelectron spectroscopy (XPS), Fourier transform IR (FTIR), RS, and/or carbon–fluorine spectroscopy (CFS), a recent innovative and unique, *in-house* developed technology [99]. Also, the molecular behaviors (e.g., electro-oxidation, conductivity, calculation of the molecular surface density, and deposition of NPs) on a graphenic surface can be carefully investigated by cyclic voltammetry (CV), square-wave voltammetry (SWV), differential pulse voltammetry (DPV), chronoamperometry (CA), and/or electrochemical impedance spectroscopy (EIS) techniques.

Eventually, immobilization techniques in the fabrication of G nanomaterial-based electrochemical biosensors were reported with special focus on the fast-emerging enzymatic biosensors [100].

13.3.2 Major Graphene-Based Biosensing Nanosystems

A growing body of evidence showed that functionalized G and G derivatives (e.g., GO, rGO, and various G- and GO-based hybrid nanocomposites) can be used as greater carbon platforms for biomolecular detection of protein nature than conventional techniques (e.g., enzyme-linked immunosorbent assay (ELISA) and western blots) because of the already described properties (e.g., optical and electronical) and sensing features (e.g., fast, sensitive, and specific detection).

13.3.2.1 Graphene-Based FETs Conducting channels G-based FETs (G-FETs) have been developed rapidly and are currently considered as an alternative for post-silicon electronics/biosensing.

Thereby, a recent pioneered study reported a flexible and high-performance G-liquid ion-gated FET aptasensor for the detection of vascular endothelial growth factor (VEGF, an angiogenesis biomarker) [101]. This aptasensor, based on polypyrrole-converted nitrogen-doped few-layer G, which was grown on Cu substrate by CVD, showed high sensitivity (i.e., limit detection as low as 100 fM), excellent reusability, mechanical bendability, and durability.

Also, a highly sensitive (down to ~2 ng ml^{-1} or 13 pM) and selective G-FET biosensor for protein detection was demonstrated using vertically oriented G sheets labeled with gold nanoparticle (AuNP)-antibody conjugates [102]. These G sheets were directly grown on the sensor electrode using a plasma-enhanced CVD method and function as the sensing channel. The protein detection was accomplished through measuring changes in the electrical signal from the FET sensor upon the antibody–antigen binding.

However, the lack of low cost, reliable, and large-scale preparation of G films limits their applications. Therefore, particular attention is being brought on nanoelectronic rGO-based FETs because of fast, facile, and substrate-independent fabrication

of G patterns that allow label-free detection of biomolecules and their movimentation (e.g., dynamic secretion of hormones from living cells) [98, 103, 104].

13.3.2.2 Graphene-Based Fluorescence Resonance Energy Transfer Several G- and GO-based FRET biosensors, including those for simultaneous multimolecular detection [91, 105], were recently developed as G-based nanomaterials displaying not only unique adsorption characteristics for biomolecules but also acting as "nanoquenchers" due to FRET. Nevertheless, in case of GO-based FRET, fine tuning of the oxidation of GO is required since it can strongly affect its fluorescence-quenching ability and binding interactions to biomolecules, such as single-stranded oligodeoxyribonucleotides (ssODNs), thereby leading to a broad range of sensitivity [106].

Thereby, a G-FRET biosensor has been shown to homogeneously, selectively, rapidly (about 5 min) and sensitively detect the lectin concanavalin A (ConA, a carbohydrate-binding protein) with a limit of detection of 0.8 nM, a linear range: 2.0×10^{-2} to 1.0 μM, using maltose-grafted aminopyrene self-assembled on the surface of G by means of π-stacking interaction, suggesting that G-FRET platform might also be a great tool for real-time monitoring of protein–carbohydrate interactions [107]. Quenched fluorescence between G and pyrene rings in presence of ConA was reverted by competitive binding of glucose that destroyed the π-stacking interaction (i.e., restoration of the fluorescence signal). A similar approach was reported to study peptide–protein interactions based on G-peptide complex [108].

Interestingly, GO-FRET has been extensively employed for protein immunosensing (e.g., dopamine (DA) [94], metalloproteinase 2 (MMP-2) [48], interleukin-5 [109], and thrombin [110]). For instance, a GO-based photoinduced charge transfer (PCT) label-free NIR fluorescent biosensor for DA was constructed [94]. In this study, the multiple noncovalent interactions between GO and DA as well as the ultrafast decay at the picosecond range of the NIR fluorescence of GO resulted in an effective self-assembly of DA molecules on the surface of GO and significant fluorescence quenching. This allowed the development of a PCT-based biosensor with direct readout of NIR fluorescence of GO for selective (i.e., quantitative recovery in the range of 98–115%), accurate (i.e., standard deviation of 2%), and sensitive detection of DA (i.e., 2 μM with detection limit of 94 nM). Further, a highly stable GO-FRET sensor was developed by covalent assembly of fluorescein isothiocyanate-labeled peptide (Pep-FITC) onto the GO surface, for rapid (i.e., within 3 h), accurate (i.e., relatively low standard deviation ≤7.03%), and highly sensitive detection (i.e., detection limit of 2.5 ng ml^{-1}) of matrix MMP-2 (a cancer biomarker) in human serum samples [48].

13.3.2.3 Graphene-Based CRET In recent years, G- and GO-based CRETs have also raised tremendous interest. Chemiluminescence is used as an exciting light source to construct universal and efficient G- or GO-based photoelectrochemical sensing platforms [111–118]. Interestingly, rGO-CdS nanocomposite highly improved the photovoltaic transfer efficiency, thereby increasing the analytical performance for photoelectrochemical immunoassays [113]. It is hypothesized that the bifunctionality of GO represented by high adsorbance of ssDNA and effective

quenching of organic dyes emission is due to its structural defects and can be reasonably utilized in a CRET system to achieve sensitive and selective detection of biomolecules (e.g., DNA and proteins) [114]. Besides, the molecular detection (e.g., DNA and proteins) by GO-CRET system involves the inhibition of peroxidatic activity of a HRP-mimicking DNAzyme by GO [119].

For instance, a reliable, sensitive, and stable G-based immunosensor for the electrochemical quantification of phosphorylated p53 on serine 15 (phospho-p53, a potential biomarker of γ-radiation exposure), has been reported [116]. The principle is based on sandwich immunoassay (i.e., ELISA-like protocol), where a sandwich immunocomplex is formed among phospho-p53 capture antibody, phospho-p53 antigen, biotinylated phospho-p53 detection antibody, and HRP-labeled streptavidin. The introduced HRP resulted in an electrocatalytic response corresponding to the reduction of H_2O_2 in the presence of thionine. Eventually, G served as sensor platform by promoting electron transfer and increasing the surface area to introduce a large amount of capture antibody, thereby increasing the analytical sensitivity. Under optimum conditions, the increase of response current was proportional to the phospho-p53 concentration in the detection range of 0.2–10 ng ml^{-1} with a detection limit of 0.1 ng ml^{-1}.

The amplified electrochemiluminescence (ECL) of quantum dots (QDs) has also been obtained by employing electrochemically rGO for green nanobiosensing of a number of biomolecules (e.g., acetylcholine and antigens) [118]. In the presence of dissolved oxygen (O_2) as coreactant, the QDs/rGO-modified electrode showed an increase of ECL intensity by 4.2 and 178.9 times using intrinsic QDs and QDs/GO-modified electrodes, respectively. ECL biosensors for choline and acetylcholine were fabricated on QDs/rGO-modified electrodes by covalent crosslinking of choline oxidase (ChO) and ChO-acetylcholinesterase (AChE), respectively. They had linear detection range and limit of detection of 10–210 and 8.8 μM, respectively, for choline and 10–250 and 4.7 μM, respectively, for acetylcholine.

13.3.2.4 Graphene-Based Metallic Surface Plasmon Resonance Owing to the high impermeability property of G to all gases and liquids [120], remarkable advances in G- and GO-based surface plasmon resonance (SPR) have been reported. These include the development of original labeled or unlabelled metallic SPR-G (i.e., coating of a G sheet above a metallic thin film, such as gold (Au), silver (Ag), titanium (TiO_2), platinium (Pt), cadmium (Cd), palladium (Pd), cobalt (Co), zinc (Zn), and aluminum (Al)) [121–123].

The use of metallic SPR-Gs has led to several potential molecular-sensing applications, such as sensitive, accurate high-throughput assessment of multiple biomolecular interactions, and clinical diagnosis based on the identification and quantification of disease biomarkers.

Thereby, based on numerical simulations and experimental studies, Au-G and Au-GO sensors were shown to exhibit excellent conductivity, as well as stable and sensitive biomolecular adsorption due to π-stacking interactions [41, 124]. These properties led to the development of innovative enzyme-free amperometric G-based Au SPR immunoassay platforms for highly sensitive detection of biomolecules

including disease biomarkers (e.g., carcinoembryonic antigen/CEA and alpha-fetoprotein/AFP) and other proteins (e.g., heparin) [124–131]. For instance, a novel approach to fabricate a label-free, rapid, sensitive, and simple amperometric immunosensor for the detection of CEA was described [126]. In this study, methylene blue, AuNPs, and CEA antibody were assembled LBL on a G-Nafion nanocomposite film-modified electrode. The immunosensor detected CEA in the range of 0.5–120 ng ml^{-1} with a limit of 0.17 ng ml^{-1}. The obtained results had good correlation with an established ELISA, which indicates a high precision and suitability for clinical diagnosis.

SPR imaging biosensors with a few G layers-on-Ag substrate were shown to achieve higher sensitivity than the conventional Au-film-based SPR imaging biosensor. Indeed, SLG deposited on 60 nm thick silver film amplified the SPR biomolecular imaging signal more than three times [132]. Thereby, GO sheet-mediated Ag enhancement, using functionalized GO sheets coupled with a signal-amplification method based on the nanomaterial-promoted reduction of Ag^+, holds a great potential for the rapid, highly sensitive, and selective analysis of biomolecules (e.g., proteinic factors such as platelet-derived growth factor (PDGF) [133] and thrombin [92]).

Several reports have demonstrated that the combination of G and PtNPs was more effective in enhancing biosensing than either nanomaterial alone. Therefore, several ways of fabricating Pt-G hybrid nanocomposites (e.g., G/PtNPs/polyaniline (PA) [134], Pt-G [135], GO-Pt-black [136], and rGO-Pt [137]) for biological sensing enhancement have been recently highlighted. For instance, in a recent study, rGOs were used as matrices to immobilize the redox probes, which were subsequently coated with PtNPs to form PtNPs-redox probes-rGOs nanocomposites [137]. With the employment of such prepared nanocomposites, a signal-amplification strategy was described based on bienzymes (i.e., GOD and HRP) modified PtNPs-redox probes-rGS nanocomposites (i.e., tracer labels for secondary aptamers through sandwiched assay). The use of AuNPs functionalized single-walled CNTs (AuNPs–SWCNTs) as the biosensor platform enhanced the surface area to capture a large amount of primary aptamers, thereby amplifying the detection response. The experimental results showed that the multilabeled PtNPs-redox probes-rGOs nanocomposites display satisfactory electrochemical redox activity and high electrocatalytic activity of PtNPs and bienzyme, which exhibited high sensitivity for the simultaneous detection of PDGF and thrombin. For PDGF, a linear range of 0.01–35 nM with a detection limit of 8 pM was obtained, while for thrombin the linear range was 0.02–45 nM and the detection limit was 11 pM.

High sensitivity measurements and rapid amperometric sensing using G-TiO_2 hybrid nanocomposites have been done for a number of molecules, such as AChE, DA, and AFP [138–140]. For instance, a high-affinity immobilization (Michaelis–Menten constant K_m: 0.22 mM) of AChE (an important neurotransmitter) onto a novel unstacked TiO_2-decorated G nanohybrid, constructed by *in situ* growth of well-dispersed TiO_2 NPs on the G sheet, enabled the reliable, stable, rapid (about 3 min), and sensitive detection of AChE inhibitors (e.g., carbaryl, an organophosphate compound). The linear detection range of the sensor was 0.001–0.015 and 0.015–2 μg ml^{-1} and the detection limit was 0.3 ng ml^{-1} [139].

Usually, the heteronanostructure of G-CdS nanocomposites facilitate the spatial separation of charge carriers, which results in an enhanced photocurrent intensity, thereby making it a promising candidate for photoelectrochemical bioapplications (i.e., detection of biomolecules) [141, 142]. G-CdS nanocomposites can be rapidly prepared in a simple one-step synthesis in an aqueous solution (e.g., LBL process) [143]. Interestingly, a sensitive electrochemical immunosensor based on CdSe-GO nanoconjugates was developed for the detection of epithelial cell adhesion molecule (EpCAM) antigen, a common marker for tumors of epithelial origin [142]. The nanoconjugates were fabricated by carboxylation of GO nanosheets (GRs), which was followed by the binding of streptavidin and amine-functionalized CdSe QDs *via* carbodiimide-coupling chemistry. Thereafter, there was an immune reaction with the biotinylated secondary antibodies. As carboxylated GO has a higher density of active sites, it allowed a large number of CdSe QDs to be immobilized onto the surface of the bionanoconjugates, which enhanced the immunosensor sensitivity. It had detection limits of 100 fg ml^{-1} and 1 pg ml^{-1} in PBS buffer and serum samples, respectively, in addition to a good selectivity, reproducibility, and long-storage stability. Therefore, CdSe-GO-based immunosensing can become a promising technique for the early detection of tumor biomarker in clinical/biological samples.

Likewise with G-based Pd biosensors, there is still a paucity of reports regarding the development of G-Co based biosensors. Nevertheless, a facile strategy to incorporate high-quality Cs-hollow CoPt bimetal alloy NPs (HCoPt) onto rGO sheet (i.e., CS-HCoPt-rGO) was successfully demonstrated for protein detection [144]. The advanced biocompatible, reliable, and stable sandwich-type electrochemical aptasensor was proposed to detect thrombin with high specificity using CS-HCoPt-rGO conjugates as secondary label. The formed conjugates provided large surface area for loading plentiful redox probe thionine, HRP, and secondary aptamer. After glutaraldehyde activation, the CS-HCoPt-rGO film greatly facilitated the capture of primary aptamer and dramatically reduced the nonspecific binding. The sensor had excellent sensitivity due to the detection of conspicuously enhanced electrochemical signal of thionine, which was amplified by Cs-HCoPt alloy nanoparticles and HRP toward the catalytic reduction of H_2O_2. The aptasensor displayed an excellent performance for thrombin with a wide linear range from 1.0×10^{-12} to 5.0×10^{-8} M and a relatively low detection limit of 3.4×10^{-13} M.

To the best of our knowledge, only rare pertinent studies have reported the development of biosensors based on G-Al nanocomposite. One recent paper described a multilayered G-Al nanopore platform for the sensitive detection of DNA and DNA–protein complexes [145]. G-Al nanolaminate membranes were formed by sequentially depositing layers of G and Al, with nanopores being formed in these membranes using an electron-beam sculpting process. The resulting G-Al_2O_3 nanopores were highly robust, which significantly exhibited lower electrical noise than nanopores in pure G. They also displayed high sensitivity to electrolyte pH at low KCl concentrations due to the high buffer capacity of Al_2O_3 and permitted electrical biasing of the embedded G electrode. In a proof of principle, the folded and unfolded transport of single DNA molecules and RecA (an essential DNA-binding repair protein)-coated DNA complexes was discerned with high temporal resolution.

This could enable nanopore integration with new G-based structures, including nanoribbons and nanogaps, for single-molecule DNA sequencing and medical diagnostic applications.

13.3.2.5 Chemical Doping of G and GO Other explorations and achievements have consisted to functionalize G and derivatives by chemical modification (e.g., nitrogen [146], poly-L-lysine [147], *N*,*N*-bis-(1-aminopropyl-3-propylimidazol salt)-3,4,9,10-perylene tetracarboxylic acid diimide (PDI) [148], methylene blue (MB) [149], polyallylamine (PAA) [150], Nafion [151], polydiacetylene (PDA) [152], PA [153], and 1-pyrenebutanoic acid succinimidyl ester (PASE) [154]).

Thereby, a recent study has reported the preparation of nitrogen (N)-doped GO electrochemical sensor by thermally annealing GO and melamine mixture [146]. This sensor was able to simultaneously determine small biomolecules such as ascorbic acid (i.e., vitamin C), DA, and uric acid. It showed a wide linear response in the concentration range from 5.0×10^{-6} to 1.3×10^{-3} M, 5.0×10^{-7} to 1.7×10^{-4} M, and 1.0×10^{-7} to 2.0×10^{-5} M for vitamin C, DA, and uric acid, respectively, while the detection limits for these biomolecules were 2.2×10^{-6}, 2.5×10^{-7}, and 4.5×10^{-8} M, respectively.

Also, a label-free electrochemical immunosensor based on G-MB nanocomposite was used to detect prostate-specific antigen (PSA, a potential biomarker of prostate cancer) from serum samples [149]. The immunosensor was constructed using a nanocomposite film of G sheets-MB-CS (G-MB-CS) as electrode material. The nanocomposite film showed high binding affinity to the electrode and was used to immobilize the antibody for PSA. This amperometric biosensor was easily developed based on the response of the voltammetric peak current to the capture of PSA induced by specific antigen–antibody (Ag–Ab) reactions. The amperometric signal decreased linearly with PSA concentration in the range of 0.05–5.00 ng ml^{-1}. A low limit of detection of 13 pg ml^{-1}, which is better in comparison to the routinely used clinical devices (0–4 ng ml^{-1}), was obtained in addition to high selectivity, good reproducibility, and stability.

Moreover, a novel label-free electrochemical aptasensor based on G-PA nanocomposites film, for DA determination on human serum samples, was also reported [153]. The resulting G-PA layer exhibited good current response for DA determination. The electrochemical aptasensor showed a linear response to DA in the range of 0.007–90 nmol l^{-1} with a lower limit of detection of 0.00198 nmol l^{-1}.

GO sheets were able to adsorb the bifunctional molecule PASE, which could then covalently interact with amines of tyrosinase-protected AuNPs (Tyr-Au), to construct highly efficient enzyme-based screen printed electrode (SPE) [154]. Tyr-Au/PASE-GO formed a biocompatible hybrid nanocomposite, which was further coated onto the working electrode surface of the SPE. The cost-effective disposable tyrosinase biosensor (Tyr-Au/PASE-GO/SPE) exhibited a rapid amperometric response (i.e., less than 6 s), a high sensitivity (i.e., detection limit: 2.4×10^{-8} M and quantitation limit: 8.2×10^{-8} M), a good affinity (K_m: 0.027 mM), and a good storage stability for monitoring the phenolic compound, namely catechol. This method showed a good linearity in the range from 8.3×10^{-8} to 2.3×10^{-5} M ($r^2=0.9980$) and might be applied as well to detect catechol oxidase, an enzyme that catalyzes the oxidation of catechol.

13.3.2.6 Biological Doping of G and GO During the last 3 years, G and GO have been also functionalized with effective and strong immobilization of biomolecules (lipids, peptides, proteins including enzymes), which enabled the development of various biological sensors for protein detection [37, 155–158]. The electronic properties of G were modulated by the adsorption of charged lipid bilayer or biomolecules (e.g., chitosan (CS), a natural biopolymer) on the surface. Biorecognition events, leading to changes in membrane/biomolecules integrity, can then be monitored electrically using an electrolyte-gated biomimetic membrane-G transistor.

For instance, rGO was used as a scaffold for immobilizing a specific biomolecule onto the substrate [37]. Thereby, rGO obtained by an original "green" and safe hydrothermal method has been used as a scaffold to adsorb/immobilize hemin (as substrate) through π–π interactions [37]. The resulting hemin-modified G nanosheet (H-G)-based electrochemical biosensor was used for the determination of L-tyrosine (Tyr, an aromatic aminoacid) levels. The H-G/GCE-based biosensor exhibited high stability and high sensitivity for Tyr detection in the broad linear range from 5×10^{-7} to 2×10^{-5} M with a detection limit of 7.5×10^{-8} M. The sensitivity of this biosensor was 133 times higher than that of bare GCE.

13.4 CONCLUSION AND PERSPECTIVES

The growing demand for compact point-of-care medical devices and portable instruments for on-site environmental sampling has stimulated an intense research on flexible sensors that can be miniaturized and function under considerable physical deformation. Also, understanding the interface between biomolecules and/or chemicals and nanomaterials is crucial for rational design and optimization of biosensors and drug delivery systems, which can be simulated by computing analysis. Interestingly, carbon nanomaterials offer unique advantages in several areas, like high surface-volume ratio, high electrical conductivity, chemical stability, and strong mechanical strength, and are thus frequently incorporated into sensing elements. Carbon nanomaterial-based sensors generally have higher sensitivities and a lower detection limit than conventional ones. With this widespread use of carbon nanomaterials in biosensors, it is timely to assess how this trend is contributing to the science and applications of biosensors. Further, the strategy of cogrowth of metal (especially Au) and metal oxides on free-standing carbon substrates is opening new possibility to develop high-performance flexible electrochemical and catalytic sensors.

Among the carbon nanoallotropes under active research, G and GO (i.e., one of the most important G forms with lattice-like nanostructure) are increasingly attracting attention for the development of advantageous biological- and chemical-sensing platforms (e.g., green and large scale nanodevices) due to their fascinating flexible structures (e.g., planar and large surface, which can be further functionalized) and physicochemical features (e.g., extraordinary mechanical, electrical, thermal, and optical properties). GO displays advantageous characteristics as a biosensing platform due to its excellent capabilities for universal highly efficient long-range quencher, direct wiring with biomolecules, heterogeneous chemical and electronic

structure, and the possibility to be processed in solution along with its ability to be tuned as insulator, semiconductor, or semimetal.

Ultimately, carbon nanomaterials in general and Gs in particular, although they still have to meet key challenges in fabrication and handling, have a bright future as simple enhanced biosensors. Especially, functionalizations of G and derivatives with diamond might enhance the overall data obtained to date. It would be very interesting for efficient translational medicine to obtain meta-analyses data to know with more accuracy which G-based nanosystem is the best for the detection of a specific protein/proteinic biomarker of disease.

CONFLICT OF INTEREST

The authors declare no conflict of interest in this work.

ABBREVIATIONS

BLG	Bilayer graphene
CEA	Carcinoembryonic antigen
CNT	Carbon nanotube
CRET	Chemiluminescence resonance energy transfer
CS	Chitosan
CVD	Chemical vapor deposition
ECL	Electrochemiluminescence
ELISA	Enzyme-linked immunosorbent assay
EpCAM	Epithelial cell adhesion molecule
ETR	Electron transfer rate
FAM	Fluorescein amidite
Fc	Ferrocene
FET	Field-effect transistor
FRET	Fluorescence/Förster resonance energy transfer
G	Graphene
GCE	Glassy carbon electrode
GO	Graphene oxide
GOD	Graphene oxidase
GP	Graphite
HRP	Horseradish peroxidase
IL	Ionic liquid
LBL	Layer-by-layer
MLG	Multi-layer graphene
MMP	Metalloproteinase
MWCNT	Multiwalled carbon nanotube
NIR	Near-infrared
NP	Nanoparticle
PEI	Polyethyleneimine
PSA	Prostate-specific antigen

QD	Quantum dots
rGO	Reduced graphene oxide
SiO_2	Silicon dioxide
SLG	Single-layer graphene
SNP	Single nucleotide polymorphism
SPE	Screen printed electrode
SWCNT	Single-walled carbon nanotube
TEM	Transmission electron microscopy

REFERENCES

[1] Boehm HP, Setton R, Stumpp E (1994) Nomenclature and terminology of graphite intercalation compounds. Pure Appl Chem, 66: 1893–1901.

[2] Boehm HP, Clauss A, Fischer GO, Hofmann U (1962a) Das Adsorptionsverhalten sehr dünner Kohlenstoffolien. Zeitschrift für anorganische und allgemeine Chemie, 316: 119–127.

[3] Boehm HP, Clauss A, Fischer G, Hofmann U (1962b) Surface properties of extremely thin graphite lamellae. Proceedings of the Fifth Conference on Carbon, Pergamon Press, London, UK; 73.

[4] Mouras S, Hamm A, Djurado D, Cousseins JC (1987) Synthesis of first stage graphite intercalation compounds with fluorides. Revue de Chimie Minerale, 24: 572–582.

[5] Geim AK, Novoselov KS (2007) The rise of graphene. Nat Mater, 6: 183–191.

[6] Geim AK, MacDonald AH (2007) Graphene: Exploring carbon flatland. Phys Today, 60: 35–41.

[7] Saito R, Fujita M, Dresselhaus G, Dresselhaus M (1992) Electronic structure of graphene tubules based on C60. Phys Rev B, 46: 1804–1811.

[8] Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Electric field effect in atomically thin carbon films. Science, 306: 666–669.

[9] Rutherford RB, Dudman RL (2002) Ultra-thin flexible expanded graphite heating element. US Patent 6,667,100 B2.

[10] Avouris P, Chen Z, Perebeinos V (2007) Carbon-based electronics. Nat Nanotechnol, 2: 605–615.

[11] Charlier JC, Eklund PC, Zhu J, Ferrari AC (2008) Electron and phonon properties of graphene: Their relationship with carbon nanotubes. In: Carbon Nanotubes: Advanced Topics in the Synthesis, Structure, Properties and Applications, Topics in Applied Physics. Jorio A, Dresselhaus G, Dresselhaus MS, Eds. Springer-Verlag, Berlin/ Heidelberg; vol. 111: 673–709.

[12] Schedin F, Geim AK, Morozov SV, Jiang D, Hill EH, Blake P, Novoselov KS (2007) Detection of individual gas molecules adsorbed on graphene. Nat Mater, 6: 652–655.

[13] Adam S, Hwang EH, Galitski VM, Das Sarma S (2007) A self-consistent theory for graphene transport. Proc Natl Acad Sci U S A, 104: 18392–18397.

[14] Chen JH, Jang C, Xiao S, Ishigami M, Fuhrer MS (2008a) Intrinsic and extrinsic performance limits of graphene devices on SiO_2. Nat Nanotechnol, 3: 206–209.

[15] Chen JH, Jang C, Williams ED, Fuhrer MS, Ishigami M (2008b) Charged impurity scattering in graphene. Nat Phys, 4: 377–381.

[16] Nair RR, Blake P, Grigorenko AN, Novoselov KS, Booth TJ, Stauber T, Peres NM, Geim AK (2008) Fine structure constant defines visual transparency of graphene. Science, 320: 1308.

[17] Kürüm U, Ekiz OÖ, Yağlıoğlu HG, Elmalı A, Ürel M, Güner H, Mızrak AK, Ortaç B, Dana A (2011) Electrochemically tunable ultrafast optical response of graphene oxide. Appl Phys Lett, 98: 1411031–1411033.

[18] Liu J, Wright AR, Zhang C, Ma Z (2008) Strong terahertz conductance of graphene nanoribbons under a magnetic field. Appl Phys Lett, 93: 041106–041110.

[19] Bao Q, Zhang H, Wang Y, Ni Z, Yan Y, Shen ZX, Loh KP, Tang DY (2009) Atomic-layer graphene as a saturable absorber for ultrafast pulsed lasers. Adv Funct Mater, 19: 3077–3083.

[20] Zhang H, Bao Q, Tang D, Zhao L, Loh K (2009) Large energy soliton erbium-doped fiber laser with a graphene-polymer composite mode locker. Appl Phys Lett, 95: P141103.

[21] Zheng Z, Zhao C, Lu S, Chen Y, Li Y, Zhang H, Wen S (2012) Microwave and optical saturable absorption in graphene. Opt Express, 20: 23201–23214.

[22] Zhang H, Virally S, Bao Q, Loh KP, Massar S, Godbout N, Kockaert P (2012) Z-scan measurement of the nonlinear refractive index of graphene. Opt Lett, 37: 1856–1858.

[23] Lee C, Wei X, Kysar JW, Hone J (2008) Measurement of the elastic properties and intrinsic strength of monolayer graphene. Science, 321: 385–388.

[24] Balandin AA, Ghosh S, Bao W, Calizo I, Teweldebrhan D, Miao F, Lau CN (2008) Superior thermal conductivity of single-layer graphene. Nano Lett, 8: 902–907.

[25] Chen S, Wu Q, Mishra C, Kang J, Zhang H, Cho K, Cai W, Balandin AA, Ruoff RS (2012) Thermal conductivity of isotopically modified graphene. Nat Mater, 11: 203–207.

[26] Saito K, Nakamura J, Natori A (2007) Ballistic thermal conductance of a graphene sheet. Phys Rev B, 76: 115409.

[27] Qizhen L, Xuxia Y, Wei W, Yan L, Ching PW (2011) A three-dimensional vertically aligned functionalized multilayer graphene architecture: An approach for graphene-based thermal interfacial materials. ACS Nano, 5: 2392–2401.

[28] Mingo N, Broido DA (2005) Carbon nanotube ballistic thermal conductance and its limits. Phys Rev Lett, 95: 096105.

[29] Heyrovska R (2008) Atomic structures of graphene, benzene and methane with bond lengths as sums of the single, double and resonance bond radii of carbon. arXiv: 0804.4086.

[30] Husale BS, Sahoo S, Radenovic A, Traversi F, Annibale P, Kis A (2010) ssDNA binding reveals the atomic structure of graphene. Langmuir, 26: 18078–18082.

[31] Segal M (2009) Selling graphene by the ton. Nat Nanotechnol, 4: 612–614.

[32] Geim A (2009) Graphene: Status and prospects. Science, 324: 1530–1534.

[33] Bae S, Kim H, Lee Y, Xu X, Park JS, Zheng Y, Balakrishnan J, Lei T, Kim HR, Song YI, Kim YJ, Kim KS, Ozyilmaz B, Ahn JH, Hong BH, Iijima S (2010) Roll-to-roll production of 30-inch graphene films for transparent electrodes. Nat Nanotechnol, 5: 574–578.

[34] Park JS, Cho SM, Kim WJ, Park J, Yoo PJ (2011) Fabrication of graphene thin films based on layer-by-layer self-assembly of functionalized graphene nanosheets. ACS Appl Mater Interfaces, 3: 360–368.

[35] Zhang B, Li Q, Cui T (2012) Ultra-sensitive suspended graphene nanocomposite cancer sensors with strong suppression of electrical noise. Biosens Bioelectron, 31: 105–109.

[36] Guo HL, Wang XF, Qian QY, Wang FB, Xia XH (2009) A green approach to the synthesis of graphene nanosheets. ACS Nano, 3: 2653–2659.

[37] Wei J, Qiu J, Li L, Ren L, Zhang X, Chaudhuri J, Wang S (2012) A reduced graphene oxide based electrochemical biosensor for tyrosine detection. Nanotechnology, 23: 335707.

[38] Riedl C, Coletti C, Iwasaki T, Zakharov AA, Starke U (2009) Quasi-free-standing epitaxial graphene on SiC obtained by hydrogen intercalation. Phys Rev Lett, 103: 246804.

[39] Sutter P (2009) Epitaxial graphene: How silicon leaves the scene. Nat Mater, 8: 171–172.

[40] Pletikosić I, Kralj M, Pervan P, Brako R, Coraux J, N'Diaye AT, Busse C, Michely T (2009) Dirac cones and minigaps for graphene on Ir(111). Phys Rev Lett, 102: 056808.

[41] Song B, Li D, Qi W, Elstner M, Fan C, Fang H (2010) Graphene on Au(111): A highly conductive material with excellent adsorption properties for high-resolution bio/nanodetection and identification. ChemPhysChem, 11: 585–589.

[42] Graphene—the perfect atomic lattice. http://www.nobelprize.org/nobel_prizes/physics/laureates/2010/illpres.html (accessed on, May 21, 2013).

[43] Brownson DA, Banks CE (2010) Graphene electrochemistry: An overview of potential applications. Analyst, 135: 2768–2778.

[44] Liu Y, Dong X, Chen P (2012) Biological and chemical sensors based on graphene materials. Chem Soc Rev, 41: 2283–2307.

[45] Singh RS, Nalla V, Chen W, Andrew TS, Ji W (2011) Laser patterning of epitaxial graphene for schottky junction photodetectors. ACS Nano, 5: 5969–5975.

[46] Mohanty N, Vikas B (2008) Graphene-based single-bacterium resolution biodevice and DNA-transistor-interfacing graphene-derivatives with nano and micro scale biocomponents. Nano Lett, 8: 4469–4476.

[47] Lorenzoni M, Brandi F, Dante S, Giugni A, Torre B (2013) Simple and effective graphene laser processing for neuron patterning application. Sci Rep, 3: 1954.

[48] Song E, Cheng D, Song Y, Jiang M, Yu J, Wang Y (2013) A graphene oxide-based FRET sensor for rapid and sensitive detection of matrix metalloproteinase 2 in human serum sample. Biosens Bioelectron, 47: 445–450.

[49] Chikkaveeraiah BV, Soldà A, Choudhary D, Maran F, Rusling JF (2012) Ultrasensitive nanostructured immunosensor for stem and carcinoma cell pluripotency gatekeeper protein NANOG. Nanomedicine (Lond), 7: 957–965.

[50] Lin J, Wei Z, Zhang H, Shao M (2013) Sensitive immunosensor for the label-free determination of tumor marker based on carbon nanotubes/mesoporous silica and graphene modified electrode. Biosens Bioelectron, 41: 342–347.

[51] Shen H, Zhang L, Liu M, Zhang Z (2012) Biomedical applications of graphene. Theranostics, 2: 283–294.

[52] Hummers WS, Offeman RE (1958) Preparation of graphitic oxide. J Am Chem Soc, 80: 1339.

[53] Sanchez VC, Jachak A, Hurt RH, Kane AB (2012) Biological interactions of graphene-family nanomaterials—An interdisciplinary review. Chem Res Toxicol, 25: 15–34.

[54] Sasidharan A, Panchakarla LS, Sadanandan AR, Ashokan A, Chandran P, Girish CM, Menon D, Nair SV, Rao CN, Koyakutty M (2012) Hemocompatibility and macrophage response of pristine and functionalized graphene. Small, 8: 1251–1263.

[55] Liu X, Sen S, Liu J, Kulaots I, Geohegan D, Kane A, Puretzky AA, Rouleau CM, More KL, Palmore GT, Hurt RH (2011) Antioxidant deactivation on graphenic nanocarbon surfaces. Small, 7: 2775–2785.

[56] Yan L, Zhao F, Li S, Hu Z, Zhao Y (2011) Low-toxic and safe nanomaterials by surface-chemical design, carbon nanotubes, fullerenes, metallofullerenes, and graphenes. Nanoscale, 3: 362–382.

[57] Liao KH, Lin YS, Macosko CW, Haynes CL (2011) Cytotoxicity of graphene oxide and graphene in human erythrocytes and skin fibroblasts. ACS Appl Mater Interfaces, 3: 2607–2615.

[58] Hartung GA, Mansoori A (2013) In vivo general trends, filtration and toxicity of nanoparticles. J Nanomater Mol Nanotechnol, 2: 3.

[59] Menaa F (2013) When pharma meets nano or the emerging era of nano-pharmaceuticals. Pharmaceut Anal Acta, 4: 223.

[60] Zan R, Ramasse QM, Bangert U, Novoselov KS (2012) Graphene reknits its holes. Nano Lett, 12: 3936–3940.

[61] Meyer J, Geim AK, Katsnelson MI, Novoselov KS, Booth TJ, Roth S (2007) The structure of suspended graphene sheets. Nature, 446: 60–63.

[62] Kasuya D, Yudasaka M, Takahashi K, Kokai F, Iijima S (2002) Selective production of single-wall carbon nanohorn aggregates and their formation mechanism. J Phys Chem B, 106: 4947.

[63] Shenderova OB, Zhirnov VV, Brenner DW (2002) Carbon nanostructures. Crit Rev Solid State Mater Sci, 27: 227.

[64] Braga S, Coluci VR, Legoas SB, Giro R, Galvão DS, Baughman RH (2004) Structure and dynamics of carbon nanoscrolls. Nano Lett, 4: 881–884.

[65] Wallace PR (1947) The band theory of graphite. Phys Rev, 71: 622–634.

[66] Semenoff GW (1984) Condensed-matter simulation of a three-dimensional anomaly. Phys Rev Lett, 53: 2449.

[67] Hernández-Ortíz S, Murguía G, Raya A (2012) Hard and soft supersymmetry breaking for 'graphinos' in uniform magnetic fields. J Phys Condens Matter, 24: 015304.

[68] Xia J, Chen F, Li J, Tao N (2009) Measurement of the quantum capacitance of graphene. Nat Nanotechnol, 4: 505–509.

[69] Chen F, Xia J, Ferry DK, Tao N (2009) Dielectric screening enhanced performance in graphene FET. Nano Lett, 9: 2571–2574.

[70] Min H; Sahu B; Banerjee SK; MacDonald AH (2007) Ab initio theory of gate induced gaps in graphene bilayers. Phys Rev B, 75: 155115.

[71] Barlas Y, Côté R, Lambert J, MacDonald AH (2010) Anomalous exciton condensation in graphene bilayers. Phys Rev Lett, 104: 096802.

[72] Min L, Brown R, Hovden P, Wojcik M, Muller DA, Park J (2012) Twinning and twisting of tri- and bilayer graphene. Nano Lett, 12: 1609.

[73] Kuzmenko AB, van Heumen E, Carbone F, van der Marel D (2008) Universal infrared conductance of graphite. Phys Rev Lett, 100: 117401.

[74] Sreekanth KV Zeng S, Shang J, Yong KT, Yu T (2012) Excitation of surface electromagnetic waves in a graphene-based Bragg grating. Sci Rep, 2: 737.

[75] Bordag M, Fialkovsky IV, Gitman DM, Vassilevich DV (2009) Casimir interaction between a perfect conductor and graphene described by the Dirac model. Phys Rev B, 80: 245406.

[76] Fialkovsky IV, Marachevskiy VN, Vassilevich DV, Marachevsky V (2011) Finite temperature Casimir effect for graphene. Phys Rev B, 84: 35446.

[77] Dobson JF, White A, Rubio A (2006) Asymptotics of the dispersion interaction: Analytic benchmarks for van der Waals energy functionals. Phys Rev Lett, 96: 073201.

[78] Artiles MS, Rout CS, Fisher TS (2011) Graphene-based hybrid materials and devices for biosensing. Adv Drug Deliv Rev, 63: 1352–1360.

[79] Wang X, Wang J, Cheng H, Yu P, Ye J, Mao L (2011) Graphene as a spacer to layer-by-layer assemble electrochemically functionalized nanostructures for molecular bioelectronic devices. Langmuir, 27(17): 11180–11186.

[80] Wujcik EK, Monty CN (2013) Nanotechnology for implantable sensors: Carbon nanotubes and graphene in medicine. Wiley Interdiscip Rev Nanomed Nanobiotechnol, 5: 233–249.

[81] Zhu Z, Garcia-Gancedo L, Flewitt AJ, Xie H, Moussy F, Milne WI (2012) A critical review of glucose biosensors based on carbon nanomaterials: Carbon nanotubes and graphene. Sensors (Basel), 12: 5996–6022.

[82] Yang W, Ratinac KR, Ringer SP, Thordarson P, Gooding JJ, Braet F (2010) Carbon nanomaterials in biosensors: Should you use nanotubes or graphene? Angew Chem Int Ed Engl, 49: 2114–2138.

[83] Chen Y, Vedala H, Kotchey GP, Audfray A, Cecioni S, Imberty A, Vidal S, Star A (2012) Electronic detection of lectins using carbohydrate-functionalized nanostructures: Graphene versus carbon nanotubes. ACS Nano, 6: 760–770.

[84] Cai Y, Li H, Du B, Yang M, Li Y, Wu D, Zhao Y, Dai Y, Wei Q (2011) Ultrasensitive electrochemical immunoassay for BRCA1 using BMIM·BF_4 coated SBA-15 as labels and functionalized graphene as enhancer. Biomaterials, 32: 2117–2123.

[85] Lu J, Liu S, Ge S, Yan M, Yu J, Hu X (2012) Ultrasensitive electrochemical immunosensor based on Au nanoparticles dotted carbon nanotube-graphene composite and functionalized mesoporous materials. Biosens Bioelectron, 33: 29–35.

[86] Kuila T, Bose S, Khanra P, Mishra AK, Kim NH, Lee JH (2011) Recent advances in graphene-based biosensors. Biosens Bioelectron, 26: 4637–4648.

[87] Wang Y, Li Z, Wang J, Li J, Lin Y (2011) Graphene and graphene oxide: Biofunctionalization and applications in biotechnology. Trends Biotechnol, 29: 205–212.

[88] Wang Z, Huang P, Bhirde A, Jin A, Ma Y, Niu G, Neamati N, Chen X (2012) A nanoscale graphene oxide-peptide biosensor for real-time specific biomarker detection on the cell surface. Chem Commun (Camb), 48: 9768–9770.

[89] Morales-Narváez E, Merkoçi A (2012) Graphene oxide as an optical biosensing platform. Adv Mater, 24: 3298–3308.

[90] Lu CH, Yang HH, Zhu CL, Chen X, Chen GN (2009) A graphene platform for sensing biomolecules. Angew Chem Int Ed Engl, 48: 4785–4787.

[91] Zhang M, Yin BC, Tan W, Ye BC (2011) A versatile graphene-based fluorescence "on/off" switch for multiplex detection of various targets. Biosens Bioelectron, 26: 3260–3265.

[92] Wang Y, Yuan R, Chai Y, Yuan Y, Bai L (2012) In situ enzymatic silver enhancement based on functionalized graphene oxide and layer-by-layer assembled gold nanoparticles for ultrasensitive detection of thrombin. Biosens Bioelectron, 38: 50–54.

[93] Wang Y, Xiao Y, Ma X, Li N, Yang X (2012) Label-free and sensitive thrombin sensing on a molecularly grafted aptamer on graphene. Chem Commun (Camb), 48: 738–740.

[94] Chen JL, Yan XP, Meng K, Wang SF (2011) Graphene oxide based photoinduced charge transfer label-free near-infrared fluorescent biosensor for dopamine. Anal Chem, 83: 8787–8793.

[95] Sun CL, Lee HH, Yang JM, Wu CC (2011) The simultaneous electrochemical detection of ascorbic acid, dopamine, and uric acid using graphene/size-selected Pt nanocomposites. Biosens Bioelectron, 26: 3450–3455.

[96] Feng T, Feng D, Shi W, Li X, Ma H (2012) A graphene oxide-peptide fluorescence sensor for proteolytically active prostate-specific antigen. Mol Biosyst, 8: 1441–1445.

[97] Nalla V, Polavarapu L, Manga KK, Goh BM, Loh KP, Xu QH, Ji W (2010) Transient photoconductivity and femtosecond nonlinear optical properties of a conjugated polymer-graphene oxide composite. Nanotechnology, 21: 415203.

[98] He Q, Sudibya HG, Yin Z, Wu S, Li H, Boey F, Huang W, Chen P, Zhang H (2010) Centimeter-long and large-scale micropatterns of reduced graphene oxide films: Fabrication and sensing applications. ACS Nano, 4: 3201–3208.

[99] Menaa F, Menaa B, Sharts O (2011) Development of carbon-fluorine spectroscopy for pharmaceutical and biomedical applications. Faraday Discuss, 149: 269–278; discussion 333–356.

[100] Putzbach W, Ronkainen NJ (2013) Immobilization techniques in the fabrication of nanomaterial-based electrochemical biosensors: A review. Sensors, 13: 4811–4840.

[101] Kwon OS, Park SJ, Hong JY, Han AR, Lee JS, Lee JS, Oh JH, Jang J (2012) Flexible FET-type VEGF aptasensor based on nitrogen-doped graphene converted from conducting polymer. ACS Nano, 6: 1486–1493.

[102] Mao S, Yu K, Chang J, Steeber DA, Ocola LE, Chen J (2013) Direct growth of vertically-oriented graphene for field-effect transistor biosensor. Sci Rep, 3: 1696.

[103] Stine R, Robinson JT, Sheehan PE, Tamanaha CR (2010) Real-time DNA detection using reduced graphene oxide field effect transistors Adv Mater, 22: 5297–5300.

[104] Dong X, Long Q, Wang J, Chan-Park MB, Huang Y, Huang W, Chen P (2011) A graphene nanoribbon network and its biosensing application. Nanoscale, 3: 5156–5160.

[105] Wu S, Duan N, Ma X, Xia Y, Wang H, Wang Z, Zhang Q (2012) Multiplexed fluorescence resonance energy transfer aptasensor between upconversion nanoparticles and graphene oxide for the simultaneous determination of mycotoxins. Anal Chem, 84: 6263–6270.

[106] Hong BJ, An Z, Compton OC, Nguyen ST (2012) Tunable biomolecular interaction and fluorescence quenching ability of graphene oxide: Application to "turn-on" DNA sensing in biological media. Small, 8: 2469–2476.

[107] Chen Q, Wei W, Lin JM (2011) Homogeneous detection of concanavalin A using pyrene-conjugated maltose assembled graphene based on fluorescence resonance energy transfer. Biosens Bioelectron, 26: 4497–4502.

[108] Lu CH, Li J, Zhang XL, Zheng AX, Yang HH, Chen X, Chen GN (2011) General approach for monitoring peptide-protein interactions based on graphene-peptide complex. Anal Chem, 83: 7276–7282.

[109] Lim SY, Ahn J, Lee JS, Kim MG, Park CB (2012) Graphene-oxide-based immunosensing through fluorescence quenching by peroxidase-catalyzed polymerization. Small, 8: 1994–1999.

[110] Chang H, Tang L, Wang Y, Jiang J, Li J (2010) Graphene fluorescence resonance energy transfer aptasensor for the thrombin detection. Anal Chem, 82: 2341–2346.

[111] Wang J, Han H, Jiang X, Huang L, Chen L, Li N (2012) Quantum dot-based near-infrared electrochemiluminescent immunosensor with gold nanoparticle-graphene-nanosheet hybrids and silica nanospheres double-assisted signal amplification. Anal Chem, 84: 4893–4899.

[112] Cheng Y, Yuan R, Chai Y, Niu H, Cao Y, Liu H, Bai L, Yuan Y (2012) Highly sensitive luminol electrochemiluminescence immunosensor based on ZnO nanoparticles and glucose oxidase decorated graphene for cancer biomarker detection. Anal Chim Acta, 745: 137–142.

[113] Tu W, Wang W, Lei J, Deng S, Ju H (2012) Chemiluminescence excited photoelectrochemistry using graphene-quantum dots nanocomposite forbiosensing. Chem Commun (Camb), 48: 6535–6537.

[114] Bi S, Zhao T, Luo B (2012) A graphene oxide platform for the assay of biomolecules based on chemiluminescence resonance energy transfer. Chem Commun (Camb), 48: 106–108.

[115] Guo Z, Hao T, Duan J, Wang S, Wei D (2012) Electrochemiluminescence immunosensor based on graphene-CdS quantum dots-agarose composite for the ultrasensitive detection of alpha fetoprotein. Talanta, 89: 27–32.

[116] Xie Y, Chen A, Du D, Lin Y (2011) Graphene-based immunosensor for electrochemical quantification of phosphorylated p53 (S15). Anal Chim Acta, 699: 44–48.

[117] Deng S, Lei J, Cheng L, Zhang Y, Ju H (2011) Amplified electrochemiluminescence of quantum dots by electrochemically reduced graphene oxide for nanobiosensing of acetylcholine. Biosens Bioelectron, 26: 4552–4558.

[118] Haque AM, Park H, Sung D, Jon S, Choi SY, Kim K (2012) An electrochemically reduced graphene oxide-based electrochemical immunosensing platform for ultrasensitive antigen detection. Anal Chem, 84: 1871–1878.

[119] Luo M, Chen X, Zhou G, Xiang X, Chen L, Ji X, He Z (2012) Chemiluminescence biosensors for DNA detection using graphene oxide and a horseradish peroxidase-mimicking DNAzyme. Chem Commun (Camb), 48: 1126–1128.

[120] Nair RR, Wu HA, Jayaram PN, Grigorieva IV; Geim AK (2012) Unimpeded permeation of water through helium-leak-tight graphene-based membranes. Science, 335: 442–444.

[121] Wang L, Zhu C, Han L, Jin L, Zhou M, Dong S (2011) Label-free, regenerative and sensitive surface plasmon resonance and electrochemical aptasensors based on graphene. Chem Commun (Camb), 47: 7794–7796.

[122] Gutés A, Carraro C, Maboudian R (2012) Single-layer CVD grown graphene decorated with metal nanoparticles as a promising biosensing platform. Biosens Bioelectron, 33: 56–59.

[123] Giovanni M, Poh HL, Ambrosi A, Zhao G, Sofer Z, Šaněk F, Khezri B, Webster RD, Pumera M (2012) Noble metal (Pd, Ru, Rh, Pt, Au, Ag) doped graphene hybrids for electrocatalysis. Nanoscale, 4: 5002–5008.

[124] Zhang B, Tang D, Liu B, Chen H, Cui Y, Chen G (2011) GoldMag nanocomposite-functionalized graphene sensing platform for one-step electrochemical immunoassay of alpha-fetoprotein. Biosens Bioelectron, 28: 174–180.

[125] Liu B, Tang D, Tang J, Su B, Li Q, Chen G (2011) A graphene-based Au(111) platform for electrochemical biosensing based catalytic recycling of products on gold nanoflowers. Analyst, 136: 2218–2220.

[126] Li Y, Yang WK, Fan MQ, Liu A (2011) A sensitive label-free amperometric CEA immunosensor based on graphene-nafion nanocomposite film as an enhanced sensing platform. Anal Sci, 27: 727.

[127] Yang M, Choi BG, Park TJ, Heo NS, Hong WH, Lee SY (2011) Site-specific immobilization of gold binding polypeptide on gold nanoparticle-coated graphene sheet for biosensor application. Nanoscale, 3: 2950–2956.

[128] Fu X, Chen L, Li J, Lin M, You H, Wang W (2012) Label-free colorimetric sensor for ultrasensitive detection of heparin based on color quenching of gold nanorods by graphene oxide. Biosens Bioelectron, 34: 227–231.

[129] Lian W, Liu S, Yu J, Xing X, Li J, Cui M, Huang J (2012) Electrochemical sensor based on gold nanoparticles fabricated molecularly imprinted polymer film at chitosan-platinum nanoparticles/graphene-gold nanoparticles double nanocomposites modified electrode for detection of erythromycin. Biosens Bioelectron, 38: 163–169.

[130] Su B, Tang D, Li Q, Tang J, Chen G (2011) Gold-silver-graphene hybrid nanosheets-based sensors for sensitive amperometric immunoassay of alpha-fetoprotein using nanogold-enclosed titania nanoparticles as labels. Anal Chim Acta, 692: 116–124.

[131] Han J, Zhuo Y, Chai YQ, Mao L, Yuan YL, Yuan R (2011) Highly conducting gold nanoparticles-graphene nanohybrid films for ultrasensitive detection of carcinoembryonic antigen. Talanta, 85: 130–135.

[132] Choi SH, Kim YL, Byun KM (2011) Graphene-on-silver substrates for sensitive surface plasmon resonance imaging biosensors. Opt Express, 19: 458–466.

[133] Qu F, Lu H, Yang M, Deng C (2011) Electrochemical immunosensor based on electron transfer mediated by graphene oxide initiated silver enhancement. Biosens Bioelectron, 26: 4810–4814.

[134] Qiu JD, Shi L, Liang RP, Wang GC, Xia XH (2012) Controllable deposition of a platinum nanoparticle ensemble on a polyaniline/graphene hybrid as a novel electrode material for electrochemical sensing. Chemistry, 18: 7950–7959.

[135] Sun W, Li L, Lei B, Li T, Ju X, Wang X, Li G, Sun Z (2013) Fabrication of graphene-platinum nanocomposite for the direct electrochemistry and electrocatalysis of myoglobin. Mater Sci Eng C Mater Biol Appl, 33: 1907–1913.

[136] Shi J, Zhang H, Snyder A, Wang MX, Xie J, Marshall Porterfield D, Stanciu LA (2012) An aqueous media based approach for the preparation of a biosensor platform composed of graphene oxide and Pt-black. Biosens Bioelectron, 38: 314–320.

[137] Bai L, Yuan R, Chai Y, Zhuo Y, Yuan Y, Wang Y (2012) Simultaneous electrochemical detection of multiple analytes based on dual signal amplification of single-walled carbon nanotubes and multi-labeled graphene sheets. Biomaterials, 33: 1090–1096.

[138] Guo S, Wen D, Zhai Y, Dong S, Wang E (2010) Platinum nanoparticle ensemble-on-graphene hybrid nanosheet: One-pot, rapid synthesis, and used as new electrode material for electrochemical sensing. ACS Nano, 4: 3959–3968.

[139] Wang K, Li HN, Wu J, Ju C, Yan JJ, Liu Q, Qiu B (2011) TiO2-decorated graphene nanohybrids for fabricating an amperometric acetylcholinesterase biosensor. Analyst, 136: 3349–3354.

[140] Huang KJ, Li J, Wu YY, Liu YM (2013) Amperometric immunobiosensor for α-fetoprotein using Au nanoparticles/chitosan/TiO(2)-graphenecomposite based platform. Bioelectrochemistry, 90: 18–23.

[141] Wang K, Liu Q, Dai L, Yan J, Ju C, Qiu B, Wu X (2011) A highly sensitive and rapid organophosphate biosensor based on enhancement of CdS-decorated graphene nanocomposite. Anal Chim Acta, 695: 84–88.

[142] Shiddiky MJ, Rauf S, Kithva PH, Trau M (2012) Graphene/quantum dot bionanoconjugates as signal amplifiers in stripping voltammetric detection of EpCAM biomarkers. Biosens Bioelectron, 35: 251–257.

[143] Zhao X, Zhou S, Jiang LP, Hou W, Shen Q, Zhu JJ (2012) Graphene-CdS nanocomposites: Facile one-step synthesis and enhanced photoelectrochemical cytosensing. Chemistry, 18: 4974–4981.

[144] Wang Y, Yuan R, Chai Y, Yuan Y, Bai L, Liao Y (2011) A multiamplification aptasensor for highly sensitive detection of thrombin based on high-quality hollow CoPtnanoparticles decorated graphene. Biosens Bioelectron, 30: 61–66.

[145] Venkatesan BM, Estrada D, Banerjee S, Jin X, Dorgan VE, Bae MH, Aluru NR, Pop E, Bashir R (2012) Stacked graphene-Al2O3 nanopore sensors for sensitive detection of DNA and DNA-protein complexes. ACS Nano, 6: 441–450.

[146] Sheng ZH, Zheng XQ, Xu JY, Bao WJ, Wang FB, Xia XH (2012) Electrochemical sensor based on nitrogen doped graphene: Simultaneous determination of ascorbic acid, dopamine and uric acid. Biosens Bioelectron, 34: 125–131.

[147] Shan C, Yang H, Han D, Zhang Q, Ivaska A, Niu L (2009) Water-soluble graphene covalently functionalized by biocompatible poly-l-lysine. Langmuir, 25: 12030–12033.

[148] Hu Y, Wang K, Zhang Q, Li F, Wu T, Niu L (2012) Decorated graphene sheets for label-free DNA impedance biosensing. Biomaterials, 33: 1097–1106.

[149] Mao K, Wu D, Li Y, Ma H, Ni Z, Yu H, Luo C, Wei Q, Du B (2012) Label-free electrochemical immunosensor based on graphene/methylene blue nanocomposite. Anal Biochem, 422: 22–27.

[150] Park S, Dikin DA, Nguyen ST, Ruoff RS (2009) Graphene oxide sheets chemically cross-linked by polyallylamine. J Phys Chem C, 113: 15801–15804.

[151] Choi BG, Park H, Park TJ, Yang MH, Kim JS, Jang SY, Heo NS, Lee SY, Kong J, Hong WH (2010) Solution chemistry of self-assembled graphene nanohybrids for high-performance flexible biosensors. ACS Nano, 4: 2910–2918.

[152] Yun JS, Yang KS, Kim do H (2011) Multifunctional polydiacetylene-graphene nanohybrids for biosensor application. J Nanosci Nanotechnol, 11: 5663–5669.

[153] Liu S, Xing X, Yu J, Lian W, Li J, Cui M, Huang J (2012) A novel label-free electrochemical aptasensor based on graphene-polyaniline composite film for dopamine determination. Biosens Bioelectron, 36: 186–191.

[154] Song W, Li DW, Li YT, Li Y, Long YT (2011) Disposable biosensor based on graphene oxide conjugated with tyrosinase assembled gold nanoparticles. Biosens Bioelectron, 26: 3181–3186.

[155] Ang PK, Jaiswal M, Lim CH, Wang Y, Sankaran J, Li A, Lim CT, Wohland T, Barbaros O, Loh KP (2010) A bioelectronic platform using a graphene-lipid bilayer interface. ACS Nano, 4: 7387–7394.

[156] Wang L, Zhang X, Xiong H, Wang S (2010) A novel nitromethane biosensor based on biocompatible conductive redox graphene-chitosan/hemoglobin/graphene/room temperature ionic liquid matrix. Biosens Bioelectron, 26: 991–995.

[157] Yue R, Lu Q, Zhou Y (2011) A novel nitrite biosensor based on single-layer graphene nanoplatelet-protein composite film. Biosens Bioelectron, 26: 4436–4441.

[158] Wu L, Feng L, Ren J, Qu X (2012) Electrochemical detection of dopamine using porphyrin-functionalized graphene. Biosens Bioelectron, 34: 57–62.

14

MODELING BRAIN DISORDERS IN SILICON NANOTECHNOLOGIES

Alice C. Parker[1], Saeid Barzegarjalali[1], Kun Yue[1], Rebecca Lee[1], and Sukanya Patil[2]

[1] *EE-Systems, University of Southern California, Los Angeles, CA, USA*
[2] *IIT Bombay, Mumbai, Maharashtra, India*

14.1 INTRODUCTION

Nanotechnologies, particularly carbon-based, could provide some future technological solutions to neural/brain disorders. This rapidly emerging research field requires close collaboration between nanotechnologists, analog circuit designers, and neuroscientists. Given this setting, we introduce some circuits modeling brain disorders. The role of nanotechnology in these models is discussed. First, we present the BioRC project. Next, we describe experiments to construct synapses with carbon nanotubes (CNTs) in the nanolab. Last, we present simulations of circuits modeling brain disorders. The research described in this chapter is at the beginning stages. The use of these simulation models to study aspects of neural disorders will be described in subsequent publications, with suggestions for such investigations provided in Section **14.20.**

Human neurons vary greatly in their properties across the brain, but most neurons in the brain contain *synapses* that connect sending neurons to receiving (postsynaptic) neurons, so that neurons can signal each other. These synapses are primarily one-directional, although some information flows back to the sending (presynaptic) neuron. The information that travels through the synapses is chemical in nature, and the chemicals are called *neurotransmitters*. Within each

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

neuron, the synapses are connected to the cell body (*soma*) of the neuron through *dendrites* that convey the electrical signals produced by the synapses. These signals are called *postsynaptic potentials* (PSPs). The entire dendritic structure is called the *dendritic arbor*. The signals can be positive (*excitatory*) or negative (*inhibitory*). The signals sum as they travel to the soma. If the resultant potential is either large enough or rising rapidly toward a large potential when the signals reach the *axon hillock*— a structure near the soma—the neuron "fires" and a spike in voltage travels down a long structure called the *axon* to reach synapses attached to recipient neurons. The structure of a basic neuron is shown in Figure 14.1. Neural connections and responses can change in strength due to learning, a property termed *plasticity*.

The synapse is one of the most important and fundamental structures of neurons, connecting neurons together. Synapses convert electrical signals in the presynaptic neuron to chemical signals that are converted again to electrical signals in the postsynaptic neuron. Synapses can be strengthened or weakened over time and can contribute to neuronal firing (excitatory synapses) or lower the possibility of neuronal firing (inhibitory synapses).

Other brain cells also contribute to cognition. There are an order of magnitude more glial cells than neurons in the brain. Oligodendrocytes are glia that form the myelin (fatty sheath) on neurons. Astrocytes are glia that contribute to processing in each neuron. These glia will be discussed later in the chapter.

This section has presented a brief introduction to the anatomy and physiology of biological neurons. A full understanding of the background required for understanding this basic neuroscience is beyond the scope of the chapter. An elementary understanding of analog electronics would be helpful as well to understand the circuits presented in the chapter.

14.2 THE BioRC PROJECT

Mead's pioneering work in neuromorphic circuits incorporating analog transistors [2] has led to circuits that are increasingly biomimetic. The Parker group in the BioRC Biomimetic Real-time Cortex project has designed neuromorphic circuits including CNT and complementary metal–oxide–semiconductor (CMOS) cortical neurons that include excitatory and inhibitory synapses, dendritic computations, and an axon hillock circuit. BioRC circuits are designed to emulate the behavior of parts of a cortical neuron. For example, the neuromorphic synapse circuit mimics the modulation of neurotransmitter quantity, receptor availability, and neurotransmitter reuptake rate, while exhibiting variations in hyperpolarizing/depolarizing potential that occur when the inhibitory/excitatory synapse responds [1, 3].

The BioRC Biomimetic Real-Time Cortex project [4] has demonstrated significant neuromorphic circuit capabilities since its inception in 2006. In collaboration with Chongwu Zhou's Nanolab, the group has constructed CNT synapses and tested them in the lab [5]. The BioRC group has designed, simulated, and fabricated a CMOS

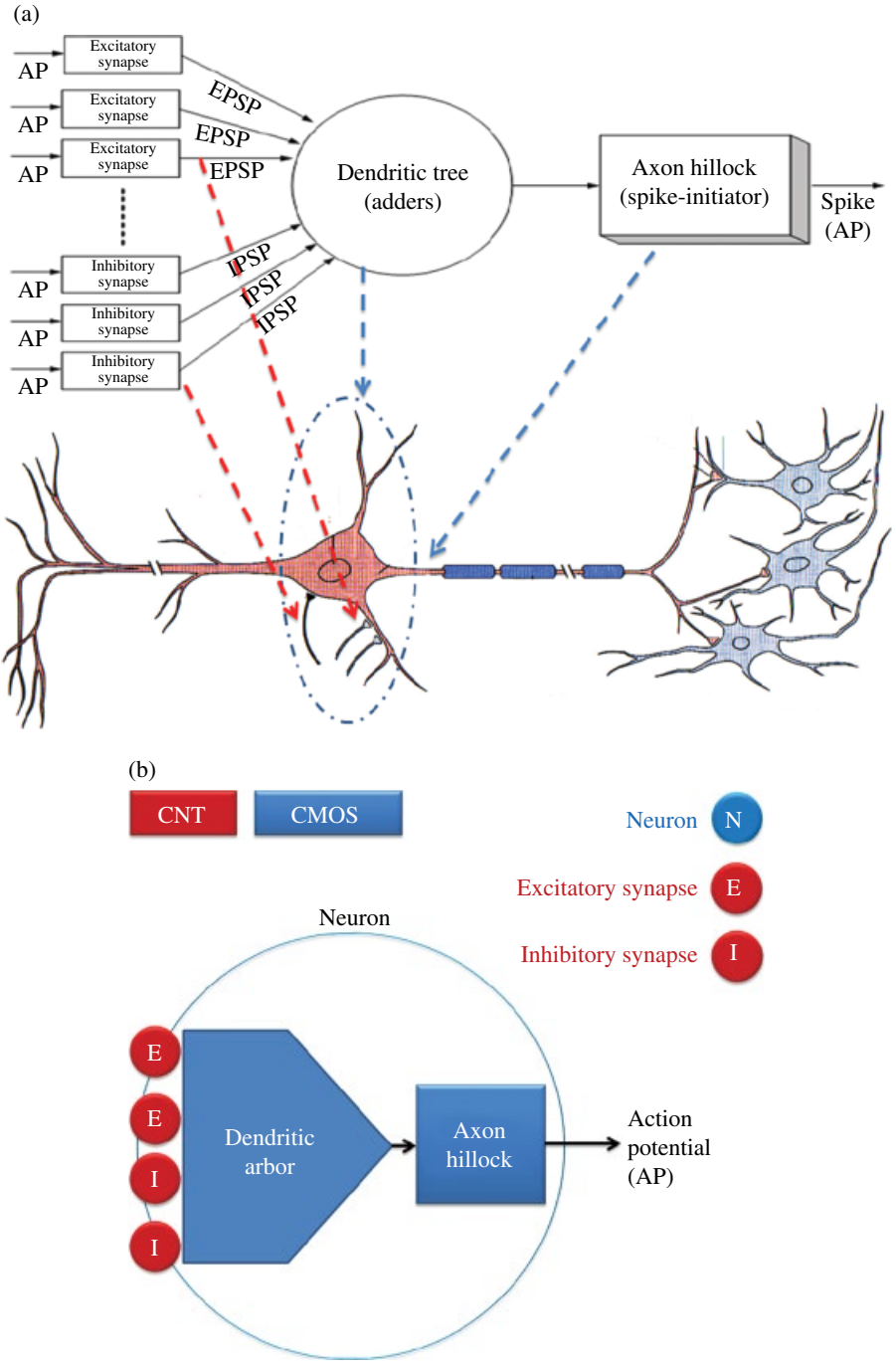


FIGURE 14.1 (a) A system block diagram of the cortical neuron model with a pyramidal neuron cartoon [1] and (b) the structure of our hybrid neuron. Reproduced with permission of IEEE.

neuron and has also simulated extensive variations of this neuron and small neural networks,[1] demonstrating the following:

- Spike-timing-dependent plasticity (STDP) of the synapse [6]
- Structural plasticity (remapping) as found in the rat barrel cortex using synapse claiming [7, 8]
- Dendritic plasticity, computations, and spiking [9, 10]
- Direction-sensitive motion detection in neuromorphic retinal starburst amacrine cells [11, 12]
- Neurons with variable behavior, both stochastic and chaotic [13, 14]
- Synchronous firing due to astrocyte slow-inward currents [15]
- Interaction between neurons and astrocytes at the tripartite synapse [16, 17].

Many of these circuits were simulated with CNT simulation models, along with CMOS versions. The group is currently researching thin-film CNT transistors to implement neuromorphic circuits, more extensive dendritic computations and dendritic spiking with plasticity, and further astrocyte–neural interactions. More extensive neural circuits with significantly more neurons and synapses are being designed. Simulation studies on circuit reliability in the face of temperature and voltage variations as well as process shifts are being performed, and the results are positive. Finally, subthreshold versions of the basic circuits are being implemented and tested. The capability of the group is at a point that extensive neural computations can be implemented with combinations of the circuits researched, with individual neural size and complexity scaled to human capabilities. The neurons can be used to form extensive networks as well.

14.3 BACKGROUND: BioRC NEURAL CIRCUITS

The most basic neuron model consist of the excitatory synapse, inhibitory synapse, dendritic arbor, and axon hillock compartments. The synapse exhibits a postsynaptic potential (PSP) when a presynaptic neuron spikes. (The PSP is positive or negative depending upon whether the synapse is excitatory or inhibitory.) In the dendritic arbor, all the PSPs from different synapses combine through nonlinear additions to output a potential that is incident on the soma. The axon hillock spikes if the summed PSP reaches a threshold. The block diagram of the spiking neuron model is shown in Figure 14.1a.

In a human cortical neuron, cell membrane potentials can vary from around −75 to 40 mV. A spike occurs when the cell potential exceeds a threshold of about 20 mV and the peak of the spike is approximately 40 mV. Our neural library includes TSMC 180 nm technology with V_{dd} of 1.8 V, TSMC 45 nm technology, and CNT field-effect

[1] This list is included to provide the reader with a sense of the range of behaviors possible with the BioRC neurons. Describing the neuroscience behind these behaviors is beyond the scope of the chapter.

transistors (FETs). Ground of the neuromorphic circuit (0.0 V) denotes the resting potential of the biological neuron and V_{dd} represents the peak of the spike (40 mV in a biological neuron). We have rescaled all the biological PSPs with respect to each technology. We have scaled the timescale for circuits (order of nanoseconds or picoseconds) compared to milliseconds in a biological neuron.[2]

We will describe the working of the spike-responsive excitatory synapse, spike-responsive inhibitory synapse, voltage adder, and axon hillock in Sections **14.3.1–14.3.4.**[3]

14.3.1 The Spike-Responsive Excitatory Synapse Circuit

The spike-responsive excitatory synapse circuit is shown in Figure 14.2. This circuit models the cell membrane potential of a postsynaptic neuron at the synapse when a spike arrives at the presynaptic neuron. The synaptic cleft potential in the electronic neuron (shown in Fig. 14.2) models the biological release of neurotransmitters stored in the presynaptic neuron into the cleft, where they bind to receptor proteins on the recipient (postsynaptic) neuron, causing the potential across the postsynaptic neural

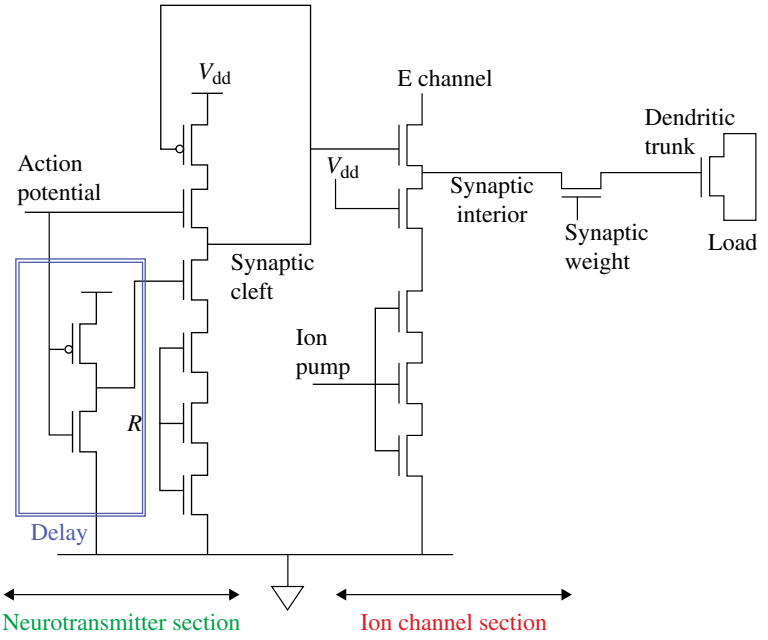


FIGURE 14.2 Schematic of the spike-responsive excitatory synapse circuit [3].

[2] The BioRC project has subthreshold neuromorphic circuits that have voltages closer to biological levels. While our neurons are quite fast compared to biological neurons, slowing the signaling in the neuromorphic circuits should not present a significant engineering challenge.

[3] We refer to the synapses as spike-responsive to differentiate them from the graded-potential synapses that we have designed for the non-spiking neurons in *Caenorhabditis elegans*.

membrane to change due to influx of ions into the postsynaptic neuron. In the BioRC neuromorphic synapse circuit, the action potential (AP) arriving at the synapse causes the voltage at the synaptic cleft to rise. This voltage is fed back to the PMOS transistor pulling up the cleft voltage to limit the rise. The increase in potential in the synaptic cleft modeling the neurotransmitter concentration will temporarily turn on the pull-up transistor connected to the ion electromotive force control, *E channel*, causing the potential at *synaptic interior* to rise. A pull-down network models the action of the return to resting potential. The control knobs like *pump* and *reuptake* tune the efficiency of the return to resting potential and neurotransmitter reuptake mechanism [3].

14.3.2 The Spike-Responsive Inhibitory Synapse Circuit

Figure 14.3 presents a spike-responsive inhibitory synapse circuit. The design is segmented into parts that facilitate biomimetic behavior corresponding to biological mechanisms. These two parts are the neurotransmitter section (synaptic cleft) and ion channel section mimicking the action of neurotransmitters and ion channels when a spike is generated in the presynaptic neuron. When an AP (spike) arrives from a presynaptic neuron, the inhibitory PSP (IPSP) signal modeled as a negative voltage occurs at the output terminal, the *dendritic trunk* [1].

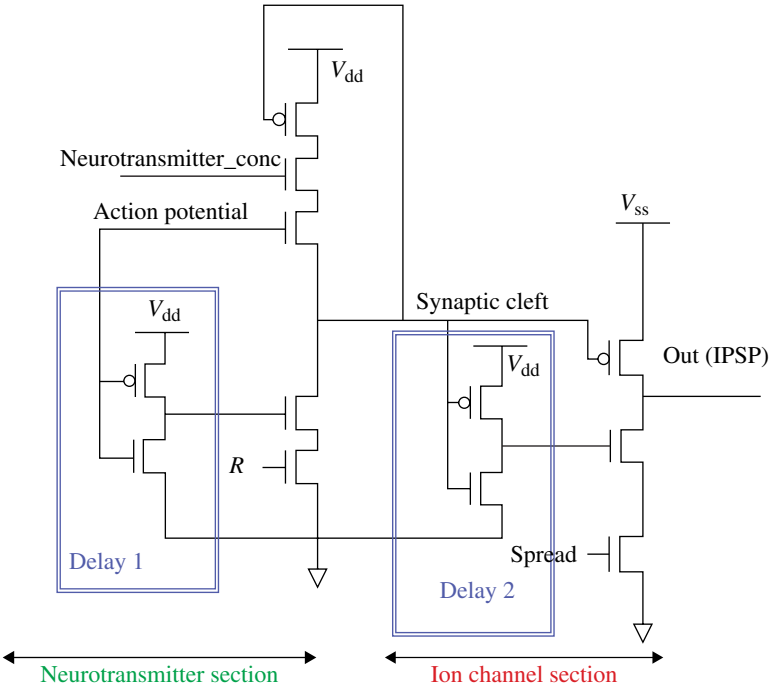


FIGURE 14.3 Schematic of the spike-responsive inhibitory synapse circuit [1]. Reproduced with permission of IEEE.

The significant difference between the spike-responsive excitatory synapse and inhibitory synapse circuits is that the second (postsynaptic) stage in the excitatory synapse is pulled up to the E channel (the potential across the membrane ion channels that is modeled as a positive voltage less than V_{dd}); while in the inhibitory synapse, the second stage is pulled to V_{ss} (a negative voltage). This makes the voltage at the output of the excitatory synapse block positive while negative for the inhibitory synapse. Both excitatory and inhibitory synapse circuit behaviors are controlled by voltages on the gates of the transistors, acting as control knobs: *reuptake, spread* (duration of the potential), and *pump*.[4]

14.3.3 Voltage Adder Circuit

Figure 14.4 is a schematic of a current-mirror adder [18]. This voltage adder performs the function of combining PSPs in the dendritic arbor. The voltage adder is used to add the excitatory PSP (EPSP) or IPSP signals from two synapses and outputs the sum of the voltage potentials as a dendritic-trunk potential. Changes to the adder can yield nonlinear additions similar to those found in dendritic structures [18].

14.3.4 Axon Hillock Circuit

In a biological neuron, if the summation of PSPs at the soma exceeds the threshold value, the axon hillock circuit (see Fig. 14.5) will generate a spike (AP) that propagates to other neurons through the axon. The circuit shown is the most basic axon hillock in the BioRC library.[5]

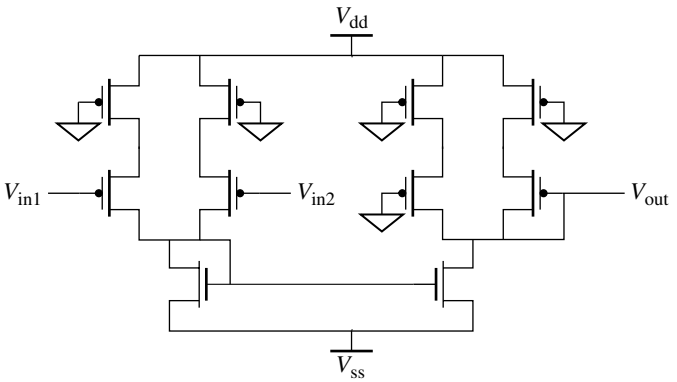


FIGURE 14.4 The voltage adder circuit [18].

[4]This inhibitory synapse is hyperpolarizing, subtracting potential, while other inhibitory synapses in the BioRC library are shunting, pulling the membrane potential down to the resting potential.

[5]The threshold in biological neurons is dynamic, responding to the rate of change of the membrane potential. Some neurons in the BioRC library possess this capability. Also neurons that spike with variable patterns and that possess spike-timing-dependent plasticity (*STDP*) are present in the library.

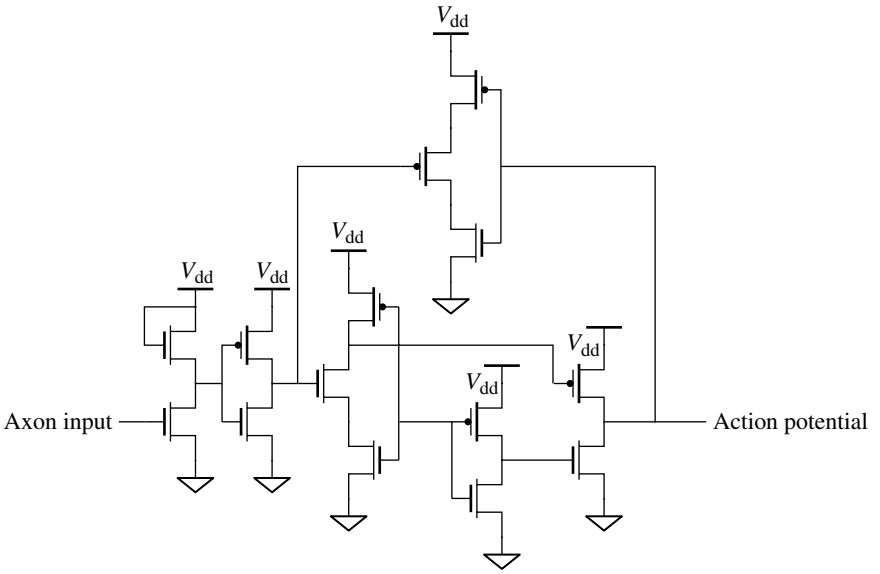


FIGURE 14.5 The axon hillock circuit.

14.4 MODELING SYNAPSES WITH CNT TRANSISTORS

Initially, for our first nanotechnology example, in collaboration with Chongwu Zhou, we constructed and tested a CNT transistor neural synapse (Fig. 14.6, [5]). This synapse used a single P-type CNT transistor with aligned nanotubes.

We then developed a neuromorphic synapse circuit that uses separated-CNT FETs. Previous work on CNT-based synapse circuits utilized aligned-CNT devices or separated CNT networks integrated with an electrochemical cell. Compared with previous work, our circuit has a lower number of transistors and is simpler to fabricate [5, 19].

14.4.1 Experimental Details for Separated CNT Transistors

A schematic of the separated CNT synapse circuit is shown in Figure 14.7. T_1 and T_2 are p-type separated-CNT FETs. The channel lengths/widths are 100/150 m for T_1 and 40/150 m for T_2. V_{REF} was held constant at −4 V. An off-chip capacitance C of 0.047 μF was used. The input is a voltage V_{AP}, an AP that impinges on the presynaptic side of a synapse. The output voltage V_{PSP} models the PSP of a biological synapse. Gate voltage V_{RG} controls the synapse strength by modulating the channel resistance of T_2 and the peak-to-peak amplitude of output V_{PSP}. The transistors were fabricated with a separated-CNT thin film as the channel material. The thin film contains a random network of semiconducting CNTs that have been separated from their metallic counterparts (Fig. 14.7). The transistors have an individual back gate structure and are built on top of an Si/SiO_2 substrate. The back gate contacts are made with Ti/Au metals. The source and drain contacts are made with Ti/Pd. The gate dielectric is Al_2O_3/SiO_2.

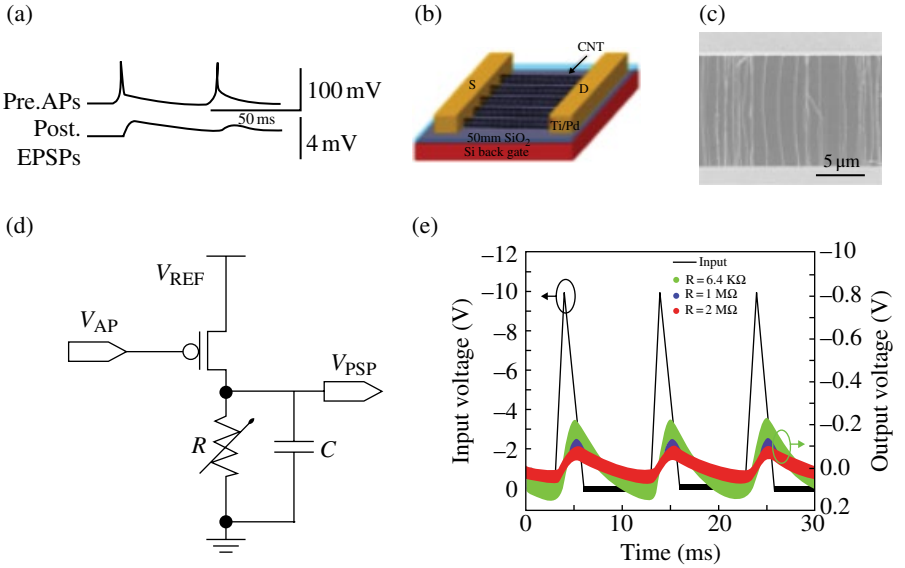
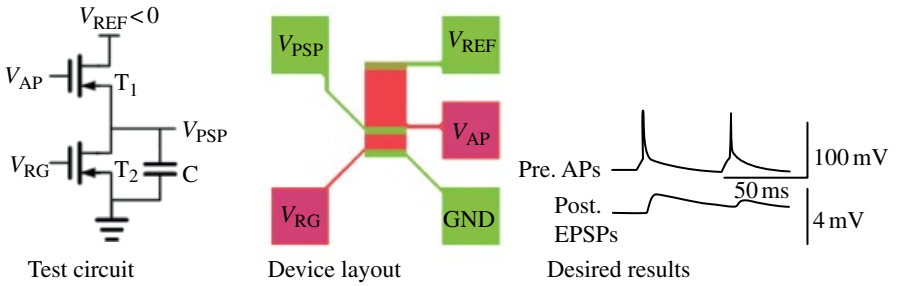


FIGURE 14.6 Synapse using carbon nanotube transistors: (a) biological action potential (AP) and resultant postsynaptic potential, (b) schematic of the nanotube transistor, (c) SEM image showing parallel nanotubes forming the transistor channel, (d) transistor-level schematic and (e) the response of the fabricated synapse to an input AP (black trace) at various resistance values. Reproduced with permission of IEEE.



- Separated CNT transistors (p-type)
- T_1 : input transistors
  - V_{AP} : Action potentail from presynaptic neuror (input signal)
- T_2 : variable synaptic strength transistor
  - V_{RG} : controls resistance from T_2
  - $V_{RG} \downarrow$ resistance $\downarrow$ synaptic strength $\uparrow$
  - $V_{RG} \uparrow$ resistance $\uparrow$ synaptic strength $\downarrow$
- $C = 0.047\ \mu F$, connected off-chip
- $V_{REF} = -4\ V$

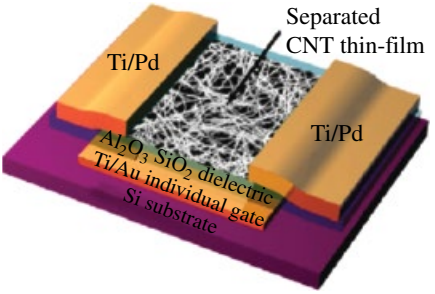
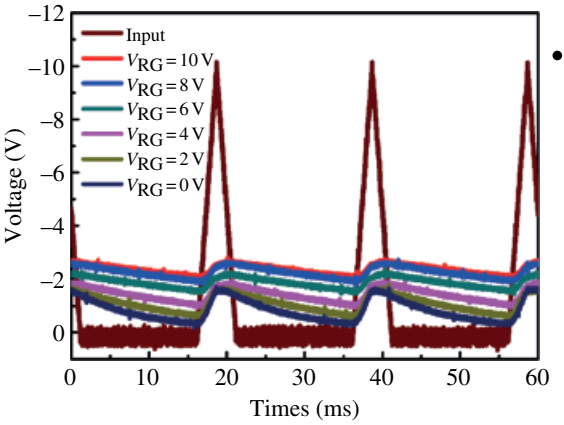


FIGURE 14.7 Thin-film carbon nanotube (CNT) transistor synapse.



- Input signal V_{AP}:
  - — Period T = 20 ms
  - — Pulse width = 5 ms
  - — Amplitude = – 10V

Output signal V_{PSP}:

- ➢ V_{RG} ↓ RC time constant ↓
- ➢ Amplitude ↓

C = 0.047 uF, V_{REF} = –4 V

FIGURE 14.8 Measured thin-film synapse signals.

14.4.2 Results for Separated CNT Transistors

Results measured from the fabricated synapse circuit are shown in Figure 14.8. The input V_{AP} was a triangular pulse with a width of 4 ms, period of 20 ms, and amplitude of −10 V. V_{RG} was varied from −10 to 0 V in 2 V steps. The peak-to-peak amplitudes ranged from 0.864 to 1.577 V and the time constants were about 19.8 ms. The amplitudes of the outputs ranged from 8.64 to 15.77% of the input amplitude, and the time constants were about four times that of the input. These results are similar to biological waveforms for excitatory PSPs. The amplitude and slope of the output decreases as V_{RG} is made more negative, tuning the strength of the synapse.

14.5 MODELING OCD WITH HYBRID CMOS/NANO CIRCUITS

The prevalence of obsessive compulsive disorder (OCD) throughout the world is thought to be between 1 and 2%, according to a range of publications, websites, and government statistics that vary depending on the population surveyed. The literature does not currently contain electronic circuit models of this disorder, but abstract models have been proposed. OCD is a debilitating neuropsychiatric disorder known for intrusive distressing thoughts (obsessions) and/or repetitive ritualized mental or behavioral acts (compulsions), normally to alleviate anxiety [20, 21]. Symptoms of OCD can be classified as follows: obsession and compulsion [20, 22, 23]. Excessive doubt about task completion, fear of contamination or uncleanliness, need for symmetry or exactness, fear of causing harm to others, excessive concern over right and wrong, and intrusive inappropriate sexual thoughts are examples of obsession symptoms. Checking, washing/cleaning, counting, ordering/arranging, repeating, hoarding, or praying are classified as compulsion symptoms.

OCD appears to involve the cortex (location of cognitive processing), the striate (an area of the cortex receiving visual signals), and the thalamus—a structure that relays signals between parts of the brain. The cortico-striatal-thalamo-cortical (CSTC) circuit in human and nonhuman primates has an essential role in controlling movement, so any damage to these areas or dysfunctional behavior of neurons or imbalance in excitatory and inhibitory synapses may cause movement disorders and even psychological problems [24]. As a case in point, blocking the activity of the sub-thalamic nucleus neurons (STNs) in human and nonhuman primates can produce disorders similar to hypokinetic and hyperkinetic disorders like Parkinson's disease (PD) and Huntington's disease [25, 26]. OCD can be explained by hyperactivity in the thalamus and cortex due to inability of the *substantia nigra pars reticulate* (SNr), a major output structure of the basal ganglia consisting of many neurons, to inhibit the thalamus [20, 27]. We have designed a biomimetic circuit that illustrates typical brain conditions and atypical conditions of OCD. Excitatory synapses and inhibitory synapses have been designed using circuits containing CNT transistors, and neuron body neuromorphic circuits have been designed using CMOS 45 nm technology transistors. Therefore, the circuit has a hybrid structure.

14.6 THE BIOLOGICAL CORTICAL NEURON AND HYBRID ELECTRONIC CORTICAL NEURON

There are approximately 100 billion neurons in the human brain. On average, every neuron receives 10,000 up to 100,000 synaptic connections from other neurons [28].[6] Realizing synapses with a nanolayer on the surface of the CMOS die is used here to enable complex neural networks to be fabricated. A SPICE simulation model from Stanford University is used for CNT transistors [29]. Due to lower parasitics and lighter weights, CNT technology is a suitable choice for the synaptic layer in neuromorphic circuits, while CMOS is used for dendritic potential adders and axon hillocks that require complicated computations. Figure 14.1a shows a biological neuron and Figure 14.1b shows our hybrid electronic neuron.

Our hybrid architecture combines inherent advantages of CMOS and CNTs. It uses CNT transistors as sensors and CMOS as complex electronic control and processing units [14], while supply voltages are identical for both technologies. Therefore, there are emerging applications for this hybrid architecture in nanotechnology. There are techniques for independent very-large-scale integration (VLSI) fabrication and CNT processing so that the CMOS part can exploit advantages of VLSI scaling like cost, flexibility, and predictable performance [30, 31].

[6]The BioRC group has designed more complicated neurons previously, which are larger compared to a neuron that is only connected to four synapses.

14.7 BIOLOGICAL OCD CIRCUIT AND BIOMIMETIC MODEL

A simplified block diagram of a mouse brain is shown in Figure 14.9a. In this diagram, several parts of the brain are shown, including multiple neurons forming each part of the brain. GPe is globus pallidus pars externalis, GPi is globus pallidus pars internalis, SNc is substantia nigra pars compacta, SNr is substantia nigra pars reticulate, and STN is sub-thalamic nucleus. Two specific types of receptors are shown, both of which are able to receive dopamine neurotransmitters. D1R is D1-type dopamine receptor and D2R is D2-type dopamine receptor. In a healthy condition, the striatum inhibits GPe, and therefore GPe stops inhibiting STN. Then, STN excites GPi/SNr and GPi/SNr inhibits the thalamus. E and I show excitatory and inhibitory synapses, respectively. Therefore, the indirect (longer) pathway on the right in Figure 14.9a acts like a brake and inhibits hyperactivity in the thalamus and cortex. On the other hand, the direct (shorter) pathway on the left can inhibit GPi/SNr, and therefore it can increase activity of the thalamus. The level of thalamus and cortical activity is controlled by this "accelerator–brake" mechanism of direct and indirect pathways [20]. Inhibitory inputs that are stimulated will help prevent the following block of neurons from firing. When the inhibitory inputs are not stimulated, the following block is more likely to fire. The excitatory inputs that are stimulated, conversely, help the following block of neurons to fire.

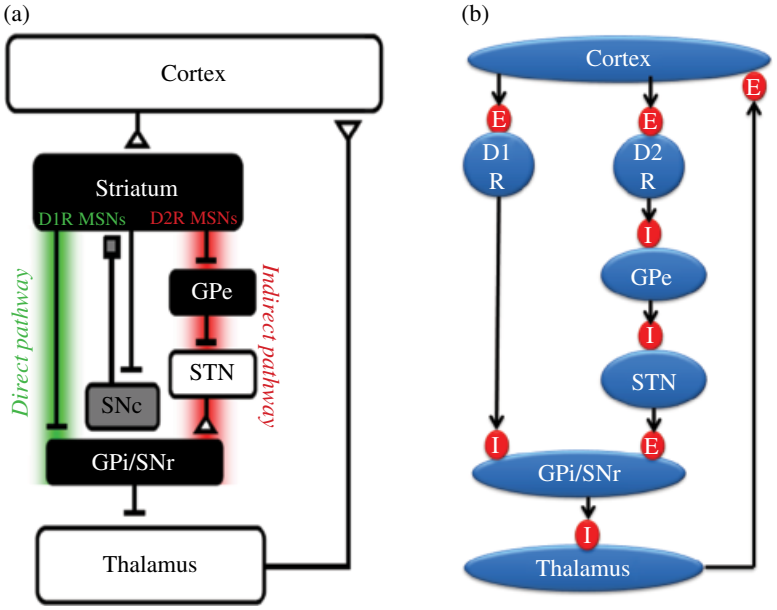


FIGURE 14.9 (a) Simplified block diagram of a circuit in the mouse brain [20]. MSN, medium spiny neuron; GPe, globus pallidus pars externalis; GPi, globus pallidus pars internalis; SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulate; STN, sub-thalamic nucleus; D1R, D1-type dopamine receptor; D2R, D2-type dopamine receptor. (b) Our biomimetic circuit mimicking part a.

Figure 14.9b shows a block diagram of our biomimetic circuit. D1 is the neuron that inhibits GPi/SNr and accelerates the activity of the *thalamus*. D2 is the neuron that excites GPi/SNr through an indirect pathway, inhibiting the *thalamus* and reducing its activity. GPi/SNr is connected to the *thalamus* through inhibitory synapses, inhibiting the *thalamus* and consequently the *cortex*. This circuit has only a few neurons and every neuron represents one part of the brain. Therefore, the circuit is a simplified version of the biological circuit. A neuromorphic circuit with more neurons can be constructed to model more detailed behavior.

Our simulations for the OCD circuit demonstrate the overall behavior of the circuit when certain neurons D1 and D2 have weak or strong synapses. We show details of the simulations in the following scenarios.

14.8 INDIRECT PATHWAY: THE BRAKING MECHANISM

Figure 14.10 shows three different firing scenarios (A, B, and C) for the indirect pathway by showing AP of all neurons.

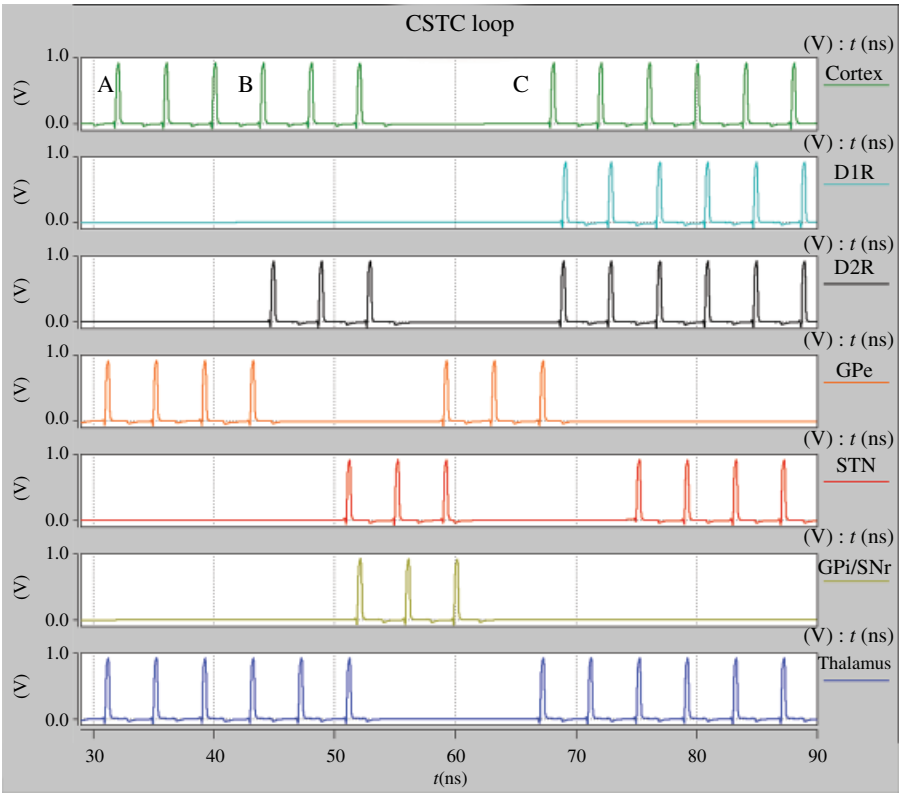


FIGURE 14.10 All neurons in the OCD circuit.

14.8.1 Indirect Pathway Is Weak

The *thalamus* keeps firing unless inhibited, and each time the *thalamus* fires it excites the *cortex*. The *cortex* tries to excite indirect pathway neuron D2, but synapses connecting the *cortex* to D2 are weak and cannot cause it to fire. Therefore, during scenario A, D2 does not fire. Since it does not fire, it cannot inhibit GPe. GPe fires unless inhibited, for it receives external excitation, which is not shown. Therefore, during scenario A, GPe keeps firing. STN fires unless inhibited, for, similar to GPe, it receives external excitation. During scenario A, GPe keeps inhibiting STN, and therefore STN does not fire. Since STN does not fire, it cannot excite GPi/SNr. GPi/SNr does not fire and consequently cannot inhibit the *thalamus*. *Thalamus* fires unless inhibited, and therefore during scenario A it shows hyperactivity.

14.8.2 Indirect Pathway Becomes Strong

Excitatory synapses connecting *cortex* to D2 get strengthened, and therefore each time the *cortex* fires, it causes D2 to fire. D2 starts inhibiting GPe. Since GPe does not fire, it stops inhibiting STN. STN fires and excites GPi/SNr. GPi/SNr inhibits the *thalamus*. The *thalamus* stops firing, and therefore it does not excite the *cortex*. The *thalamus* and *cortex* do not fire for some time. Since the *cortex* does not fire, it cannot cause D2 to fire. D2 does not fire and GPe starts firing, and STN stops firing. This means that GPi/SNr will not inhibit the *thalamus*, the *thalamus* will fire again and will excite the *cortex*. So, even if scenario B lasts for some time, the *thalamus* and *cortex* will show activity, but this feedback mechanism reduces the hyperactivity and slows down the firing rate of the *thalamus* and *cortex*.

14.8.3 Indirect Pathway Is Still Strong

In this scenario, the indirect pathway is still strong, but the direct pathway also becomes stronger and D1 starts firing each time the *cortex* fires. So, the *cortex* inhibits GPi/SNr. GPi/SNr cannot fire even though it receives excitation from brake mechanism pathway. Therefore, during scenario C, D2 and STN keep firing. However, GPi/SNr does not fire and the *thalamus* shows hyperactivity.

14.9 DIRECT PATHWAY: THE ACCELERATOR

The direct pathway behavior is also shown in Figure 14.10.

14.9.1 Accelerator Is Weak

During scenario A, neither D2 nor D1 are active. Therefore, GPi/SNr does not fire and does not inhibit *thalamus*.

14.9.2 Accelerator Is Still Weak

In scenario B, the accelerator pathway is still weak, but the brake becomes stronger by increasing the amplitude of PSP output of excitatory synapses that are connecting the *cortex* to D2, and therefore GPi/SNr starts firing, inhibiting the *thalamus*.

14.9.3 Accelerator Becomes Stronger

In scenario C, the accelerator pathway becomes stronger and starts inhibiting GPi/SNr. GPi/SNr stops inhibiting the *thalamus*, and the *thalamus* starts firing.

14.10 TYPICAL AND ATYPICAL RESPONSES

Figure 14.11 shows detailed results of the biomimetic neural circuit simulation, with synaptic responses of D1 and D2. We have divided the results into three scenarios: A, B, and C. *Thalamus*, GPe, and STN always keep firing (they are excited by external stimuli) unless they are inhibited. PSP outputs of excitatory synapses that are connecting *cortex* to D1 and D2 are shown in Figure 14.11.

14.10.1 Scenario A: *Thalamus* Keeps Firing, Neither D1 Nor D2 Are Active

In scenario A, synapses that are connecting *cortex* to both sides of striatum (D1 and D2) are weak. Therefore, they do not fire. D2 does not fire, so it cannot inhibit GPe. Therefore, GPe keeps firing and inhibiting STN. STN does not fire and cannot excite GPi/SNr. GPi/SNr does not get excited, and therefore it cannot inhibit *thalamus*. *Thalamus* keeps firing and exciting *cortex*. This means that brake pathway of the neural circuit does not work and the circuit shows hyperactivity in the *thalamus* and *cortex*.

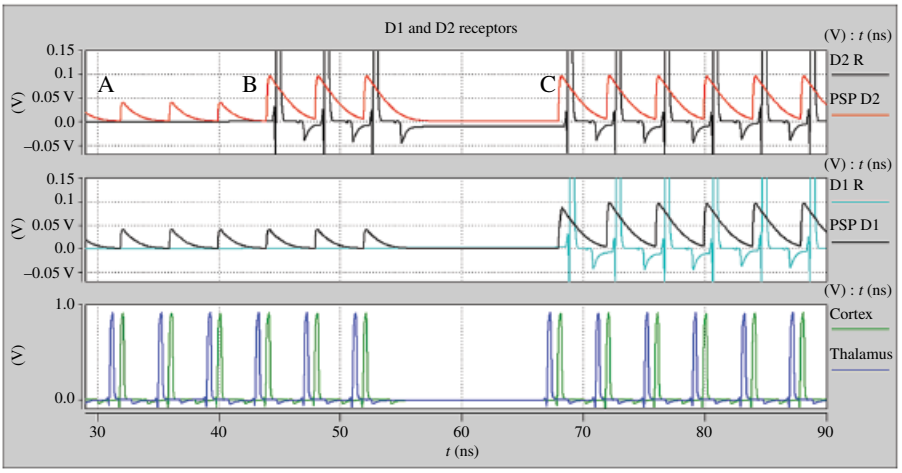


FIGURE 14.11 Indirect and direct pathways in the OCD circuit.

14.10.2 Scenario B: D2 Brakes the *Thalamus*

In scenario B, synapses that are exciting D2 by *cortex* become stronger and D2 starts firing and inhibiting GPe. Therefore, GPe stops inhibiting STN. STN starts exciting GPi/SNr. Finally, GPi/SNr starts inhibiting *thalamus* and as a consequence *thalamus* and *cortex* stop showing hyperactivity. Response of the circuit is not immediate because this braking occurs through an indirect mechanism. By having a stronger brake pathway, the neural circuit can control hyperactivity and exits the abnormal condition.

14.10.3 Scenario C: D1 Accelerates the *Thalamus*

In scenario C, D1 also becomes stronger by increasing the amplitude of excitatory synapses that are connecting *cortex* to D1 and starts inhibiting GPi/SNr and therefore accelerates activity of the *thalamus* and *cortex*. Once again, the circuit goes to OCD condition. The *Thalamus* and *cortex* show hyperactivity.

14.11 MODELING SCHIZOPHRENIC HALLUCINATIONS WITH HYBRID CMOS/NANO CIRCUITS

Recent research shows that neural processes in the human brain are highly predictive and proactive [32]. These predictions are essential, for they help us to show fast and targeted responses to environmental stimuli [33]. Hallucination can be caused by failure in prediction or erroneous prediction in schizophrenic patients. In the normal condition, the brain suppresses responses to familiar and expected input [34] and detects when unexpected inputs occur [33].

In schizophrenia, micro and macro neural circuitry of the limbic system are altered and cause schizophrenic hallucinations. This alteration can cause disinhibition or erroneous inhibition by changing the balance between excitatory and inhibitory synapses [5].

Here, a biomimetic neural circuit is designed that mimics normal and schizophrenic conditions and their responses to expected and unexpected input stimuli. This circuit has a hybrid structure. All the synapses, excitatory and inhibitory, are modeled with CNT transistors. These synapses previously had been designed and simulated by our research group [1]. Deng and Wong have developed a simulation model for CNT that is used in this work [29]. Each neuron has a two-stage dendritic adder and one axon hillock. These components are designed with transistors based on TSMC 45 nm CMOS technology. Every neuron receives a summed PSP output of four synapses, and, based on that, decides whether to fire or not. The output of the neuron is an AP. It is more likely that a neuron fires if it receives more excitatory inputs, and it is less likely that a neuron fires if it receives more inhibitory inputs.

14.12 EXPLANATION FOR SCHIZOPHRENIA SYMPTOMS

Symptoms of schizophrenia are classified as positive and negative. The lack of executive functions like planning and working memory is one of the main negative symptoms. These activities in normal conditions require an active prefrontal area in

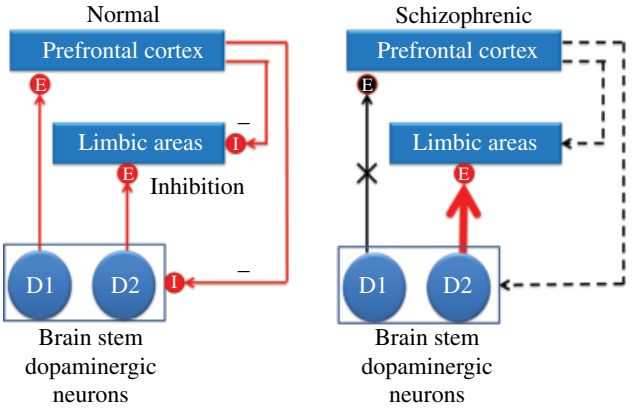


FIGURE 14.12 Neuroanatomical model of schizophrenia. Adapted from Refs. [9, 10].

the brain, so the negative symptoms of schizophrenia is due to lack of activity in the prefrontal area. Hallucinations and delusions are related to hyperactivity in the limbic area, which are important positive symptoms. Daniel Weinberger suggested that, in schizophrenia, two dopaminergic systems (neurons receiving dopamine at their synapses) are distributed in different ways, as shown in Figure 14.12 [27].

14.12.1 Normal Condition

Figure 14.12 shows that in the normal condition, the prefrontal cortex is getting stimulated by one of the dopaminergic neurons (D1). The prefrontal cortex shows activity and inhibits the limbic area. In other words, D1, through a feedback system, inhibits the limbic area. Therefore, in the normal condition, the prefrontal cortex is operating normally and the person does not present any negative symptoms. Simultaneously, limbic areas are inhibited by the prefrontal cortex (and weakly excited by D2), and therefore do not trigger hyperactivity and the person does not present positive symptoms of schizophrenia.

14.12.2 Abnormalities in the Dopaminergic System Cause Schizophrenic Symptoms

Schizophrenic symptoms are due to abnormalities in dopaminergic system. If D1 stops exciting the prefrontal cortex, then the prefrontal cortex will show a lack of activity and negative symptoms of schizophrenia. As a consequence, the prefrontal cortex will stop inhibiting the limbic area. On the other side, it will cease inhibiting D2, which is exciting the limbic area. Therefore, indirectly the prefrontal cortex will excite the limbic area. An imbalance between dopaminergic neurons can cause limbic areas to receive more excitation and less inhibition and show hyperactivity. Similarly, we have demonstrated in this chapter that neural alteration can cause symptoms of schizophrenia and supports the theory that hallucinations and other positive symptoms of schizophrenia are caused by disinhibition due to a neural alteration.

14.13 DISINHIBITION DUE TO MISWIRING

Miswiring in neural circuitry can cause disinhibition in pyramidal neurons and therefore can cause hallucination [35]. As Figure 14.13 shows, in the normal condition a simplified pyramidal neuron receives one pair of excitatory synapses and two inhibitory synapses coming from GABA interneurons. In a schizophrenic condition, there might be two pairs of excitatory synapses and only one inhibitory GABA interneuron. This means that disinhibition in this pyramidal neuron is more likely to happen in a schizophrenic condition. The GABA interneurons are simple neurons that release the neurotransmitter GABA, which makes the recipient neuron less likely to fire.

In our simulation experiments, we change the synaptic weights to mimic this structural change. We demonstrate how this change causes failure in recognizing expected patterns.

14.14 OUR HYBRID NEUROMORPHIC PREDICTION NETWORK

We have designed a simple neuromorphic prediction network that expects two spikes arriving at its input in sequence, separated by a time period that is dependent on the neural delays in the network. If the two spikes arrive as predicted, the pyramidal neuron will not fire. Spikes separated by a shorter or longer time period that does not match the predicted delay or a missing second spike will cause the final pyramidal neuron in the network to fire.

Figure 14.14 shows a block diagram of our hybrid neural circuit. Every neuron needs coincident PSP of two excitatory synapses to fire. Therefore, any spike

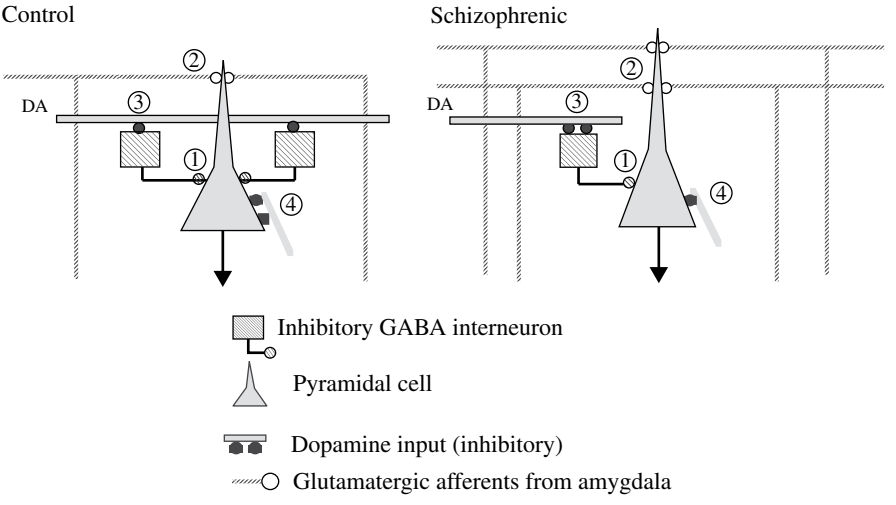


FIGURE 14.13 Miswiring in schizophrenia [35].

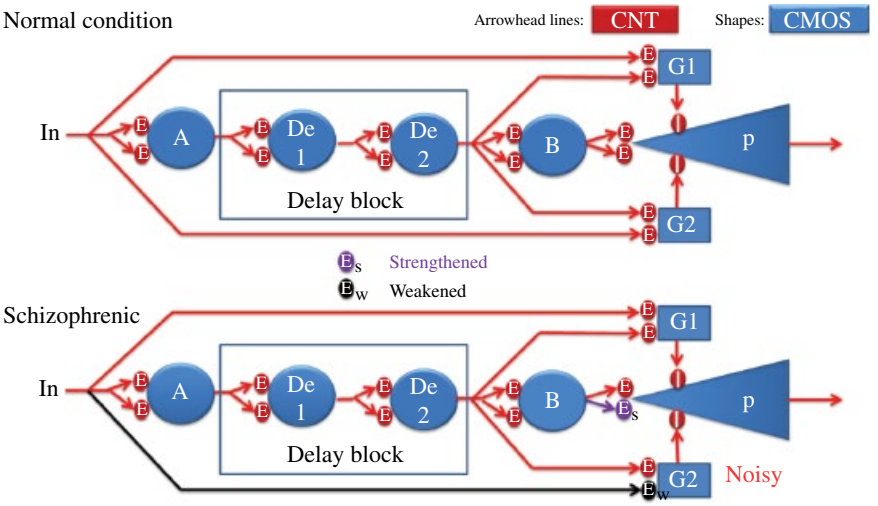


FIGURE 14.14 Block diagram of our hybrid circuit.

coming from input signal that is a triggering signal can cause neuron *A* to fire, because the input signal is connected to neuron *A* through two excitatory synapses. If *A* fires, it will cause De 1 to fire and as a consequence will cause De 2 to fire. Then De 2 will cause *B* to fire. Finally, *B* will cause a pyramidal neuron (P) to fire, unless P is inhibited by both interneurons G1 and G2. Interneurons G1 and G2 are connected to *IN* through only one excitatory synapse. This means that if the input is only one spike followed by a sufficient delay then G1 and G2 will not fire and will not inhibit P from firing.

Therefore, if the input is only one spike and the second spike does not appear at the expected time, neuron P will fire. AP output of neuron P is considered to be the final output of the circuit. The circuit is expecting two consecutive spikes; therefore if the input is only one spike, which is not predicted, the final pyramidal neuron will fire. If there are two consecutive spikes coming as the input, then when the first spike reaches one of the synapses of neurons G1 and G2, the second spike arrives at the other synapse, and therefore interneurons G1 and G2 fire and inhibit neuron P from firing. When G neurons fire, it means the circuit has detected the expected pattern (two consecutive spikes), and therefore they inhibit P.

Neurons De 1 and De 2 are delay blocks. They delay response of the circuit after the first spike and wait to see if there is another spike.

14.15 SIMULATION RESULTS

Figure 14.15 shows the simulation result for both normal and schizophrenic conditions for four different scenarios.

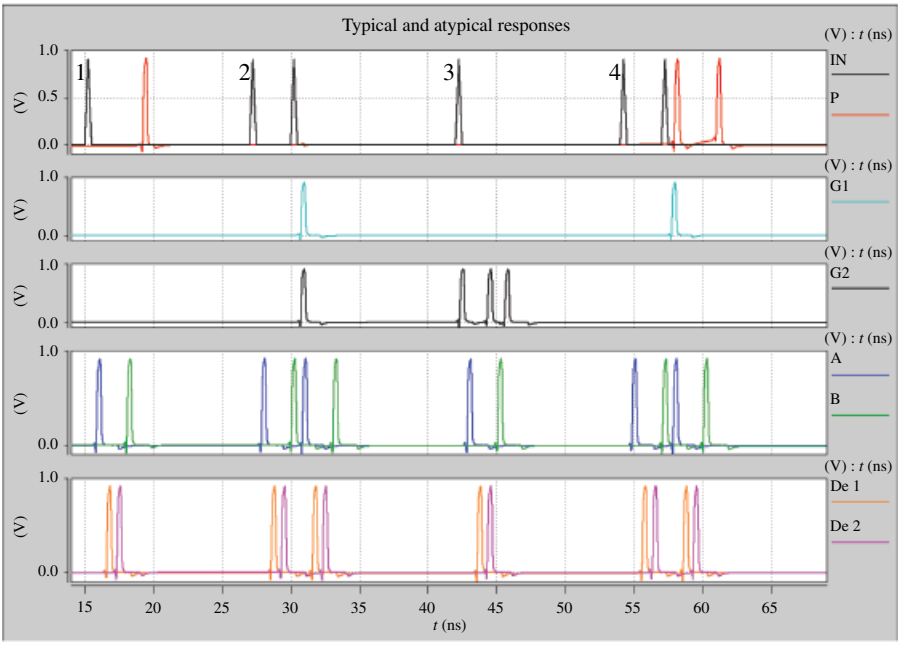


FIGURE 14.15 Simulation results for normal and schizophrenic conditions.

14.15.1 Input Scenario 1

Circuit is in normal condition and waiting for the input. The input is just one spike. It causes *A* to fire. *A* causes De 1 to fire. De 1 causes De 2 to fire. When De 2 fires, IN is silent and, therefore, G neurons do not fire. G neurons only fire if both synapses are excited simultaneously. They have not detected the expected consecutive two spikes, so they do not fire. De 2 causes *B* to fire, and finally *B* causes P to fire. It is similar to the condition when a person hears something that he does not expect and he understands something is wrong.

14.15.2 Input Scenario 2

Circuit is still in the normal condition. The input is two consecutive spikes. The first spike causes *A* to fire. *A* causes De 1 to fire. De 1 causes De 2 to fire. When De 2 is trying to cause *B* to fire, the second spike from the input arrives to the other synapse of G neurons. The *B* neuron is trying to cause P to fire, but two interneurons G1 and G2 are inhibiting P from firing. As a result, P does not fire. Therefore, when G neurons fire, it means the input is two consecutive spikes, and when P fires it means the input is not expected. This is similar to the condition when a person hears something that is consistent with his predication, and therefore the person stops responding to it.

14.15.3 Input Scenario 3

The circuit now is altered to a schizophrenic condition by altering synaptic strengths as shown in Figure 14.14. G2 has four inputs and two inputs are connected to a Gaussian noise source with zero mean and 40 mV variance as shown in Figure 14.16. This noise can cause G2 to fire when the input is only one spike or even when there is no input. This means that due to noise, the circuit mistakenly thinks the input is two consecutive spikes. Therefore, it is analogous to hallucination and mistaken cognition.

14.15.4 Input Scenario 4

The circuit is still altered to a schizophrenic condition, but the noise has stopped, for we only want to see the effect of the alteration of synaptic strengths. As Figure 14.14 shows, in a schizophrenic condition, one of the excitatory synapses going from *B* to P is strengthened. Therefore, P receives more excitation. Also, one of the excitatory synapses connecting input to G2 is weakened. Therefore, if the input is two consecutive spikes, G2 does not fire and P receives less inhibition. Receiving more excitation and less inhibition causes neuron P to fire even when the input is the expected two consecutive spikes. This disinhibition caused by neural circuitry alteration causes the network to fail in recognizing an expected pattern and displaying an appropriate response. Although the input is two consecutive spikes, G2 does not fire and P fires twice.

14.16 NUMERICAL ANALYSIS OF FALSE FIRING

Noise is common in neural circuits in the biological brain. Variations can be chaotic or stochastic (random). The purpose of using injected noise in the simulations is to show how hallucinations can occur if neurons fire unexpectedly due to noise. Neurons G1 and G2 should fire if they receive input from two synapses simultaneously. As stated previously, false is firing due to noise mimics hallucination. Figure 14.16

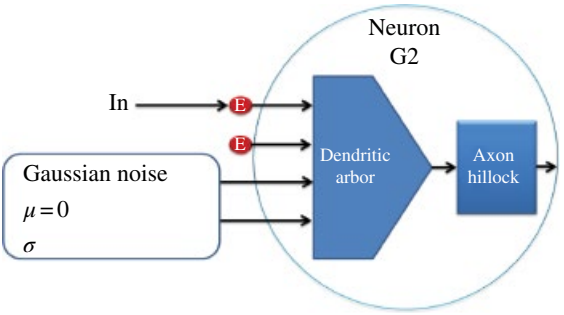


FIGURE 14.16 Noisy neuron.

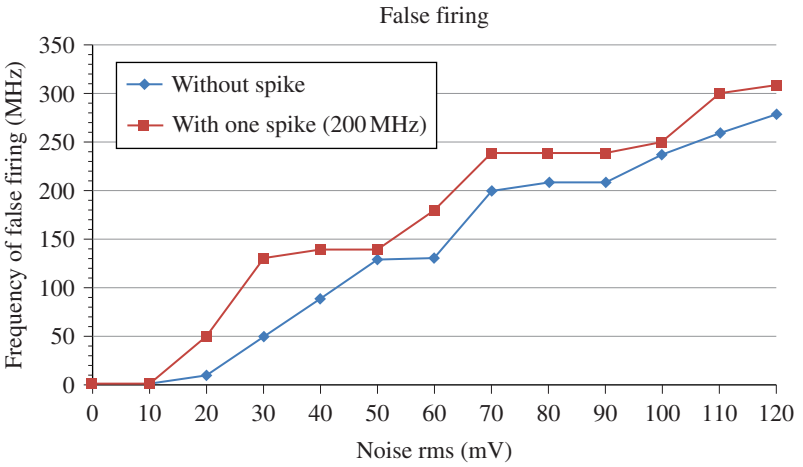


FIGURE 14.17 False firing due to noisy inputs.

shows neuron G2 with two synapses connected to a noise source. The noise source has Gaussian output with zero mean and different variances. The simulation result is shown in Figure 14.17. As the lower curve illustrates, when there is no noise or when noise has only 10 mV rms value, there is no false firing, but when the magnitude of noise increases, the frequency of false firing increases. The upper curve shows a similar simulation but this time one of the inputs to G2 receives one spike with frequency equal to 200 MHz, which makes it more likely to fire mistakenly.

14.17 MODELING PD WITH CMOS CIRCUITS

Neurological disease, including PD, affects as many as one billion people worldwide, estimated in 2006 by WHO. Modern treatments are effective at managing the early motor symptoms of PD, but the drugs eventually become ineffective at treating the symptoms. However, advancing technology, such as deep brain stimulation, has been used to reduce the motor symptoms when drugs are ineffective. This shows the significant potential of electrical technology using in neurological disorders.

We demonstrate neuromorphic circuit models in Figure 14.18 that exhibit dopamine effects on a synapse of a degenerated dopaminergic neuron exhibiting PD [36]. A neuromorphic excitatory synapse circuit is implemented to mimic a biological synapse. A dopamine-related sub-circuit is embedded in the synapse circuit to accomplish the modulatory effect of dopamine. We represent the release of dopamine with the voltage of a gate control signal. With the dose of dopamine decreasing in the synapse, the gate voltage keeps dropping until it shuts off the NMOS transistor, disabling the synapse. Meanwhile, that transistor being shut off will trigger a compensatory circuit to mimic the increase of the dopamine D2 receptors (DA-D2R) due to the plasticity.

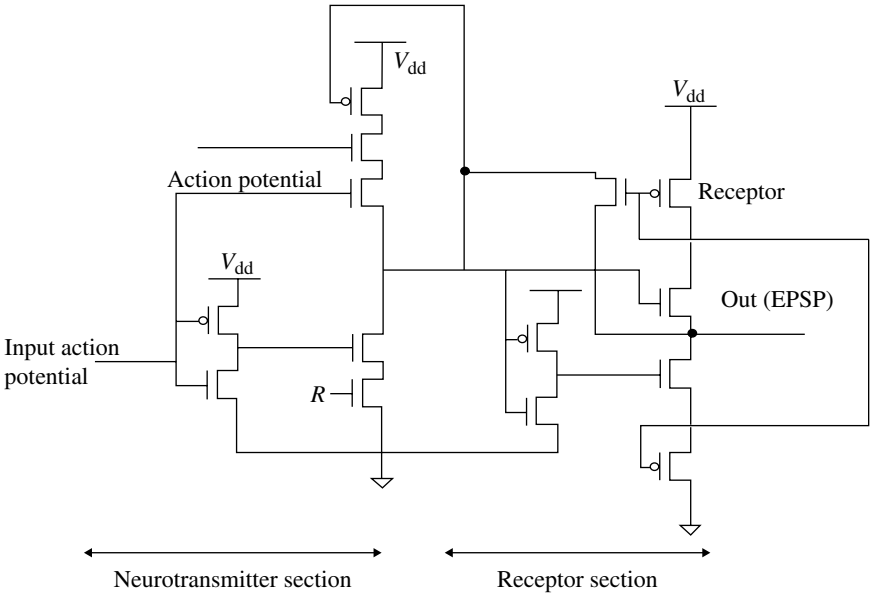


FIGURE 14.18 Dopamine circuit.

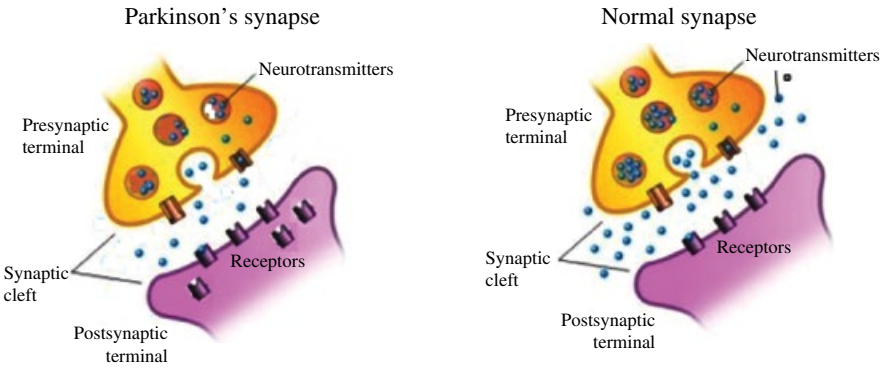


FIGURE 14.19 Degeneration in Parkinson's disease.

In PD, the loss of the dopamine D2 receptors (DA-D2R) is reportedly due to the degeneration of dopaminergic neurons, whereas the increase in DA-D2Rs results from increased expression on remaining dopaminergic terminals and/or increased synthesis within striatopallidal neurons or cholinergic interneurons (Fig. 14.19). The loss of DA-D2Rs induces the disconnection between presynapse and postsynapse; thus, we shut off the presynaptic circuit to mimic the disconnection. As the presynaptic circuit shuts down the synapse, the compensatory mechanism starts to generate more DA-D2Rs to maintain the connection. We use a compensatory circuit

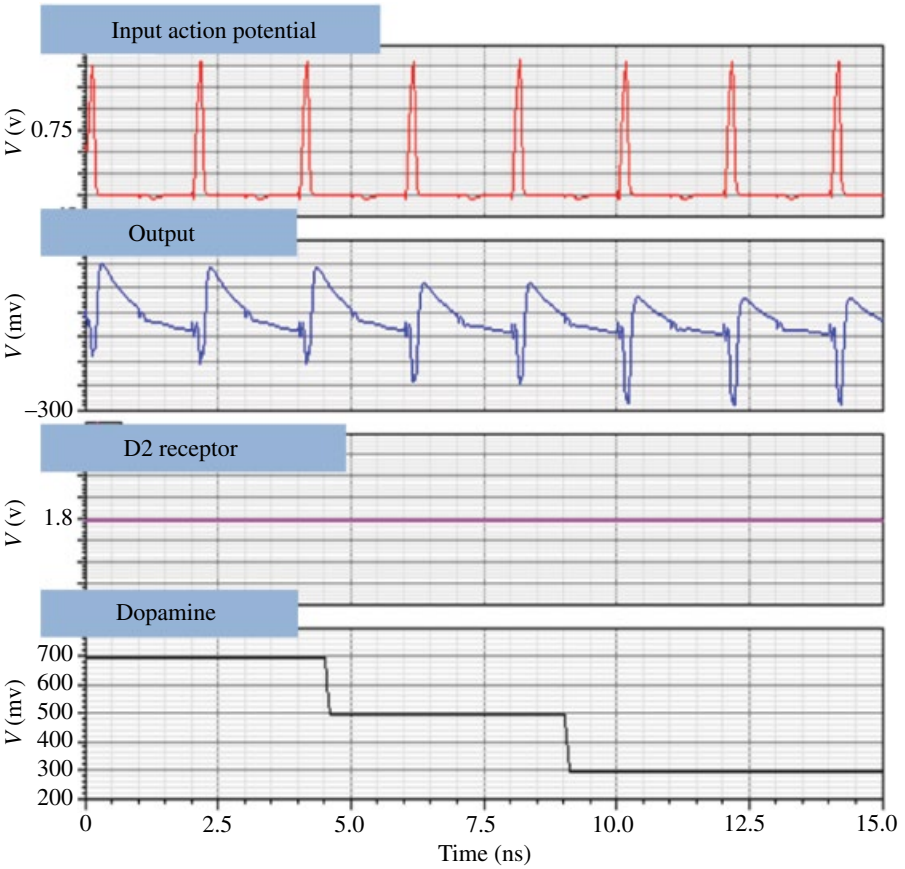


FIGURE 14.20 Simulation result of dopamine circuit.

to allow subthreshold voltage conduction. In a circuit including multiple synapses, if there is enough subthreshold voltage, the connection can still exist (Fig. 14.18).

Figure 14.20 shows a part of the simulation result. The dopamine signal keeps dropping 28.5% every 4 ns, but the EPSP only drops 14% due to the receptor signal maintaining a high voltage.

14.18 MODELING MS WITH CMOS CIRCUITS

Demyelinating disease (loss of the myelin sheath on neural axons, weakening signal transmission), such as multiple sclerosis (MS), affects over 350,000 people in the United States [37]. There is no known cure for MS and there are few therapies. However, nanotechnology may play a role for demyelinating disease treatment. The physical property of nanomaterial can model demyelinated axons. Many biocompatible nanomaterials can be fabricated as resistors and capacitors, which can be attached to a demyelinated axon

to alter the resistance and capacitance. A circuit simulation model shows the possibility of modeling healthy as well as demyelinated axons. In the future, a nanoscale neuromorphic demyelination model can be used to build a complex circuit that is close to biological brain circuits for advance disease research.

14.19 DEMYELINATION CIRCUIT

Two types of sub-modules are included in our basic neuron model: the axon hillock and the axon. In the axon hillock circuit (Fig. 14.21), the input presynaptic membrane potential initiates AP spikes. This circuit behaves in a similar fashion to a self-resetting CMOS circuit, receiving a rising edge and producing a pulse whose width is controlled by the gate delay of the inverter (transistor X8 and X7) [38].

The axon circuit is shown in Figure 14.22. The biological axon has high resistivity ($\rho = 40\ \Omega \cdot m^2$) and low capacitance ($C = 5 \times 10^{-5}\ F/m^2$). The demyelination decreases the resistivity ($\rho = 0.2\ \Omega \cdot m^2$) and increases the capacitance ($C = 10^{-5}\ F/m^2$) of the axon. This lesion of demyelination weakens the signal transmission along the axon and strengthens the leakage across the membrane. The length of every myelin sheath is set as 1 mm and the radius is 5 μm. So the resistance in the circuit model $R_m = 1.3 \times 10^9\ \Omega$, and the capacitance $C_m = 1.57 \times 10^{-12}$ F in the normal axon, and $R_m = 6.4 \times 10^6\ \Omega$, $C = 5 \times 10^{-5}\ F/m^2$ in demylinated axons.

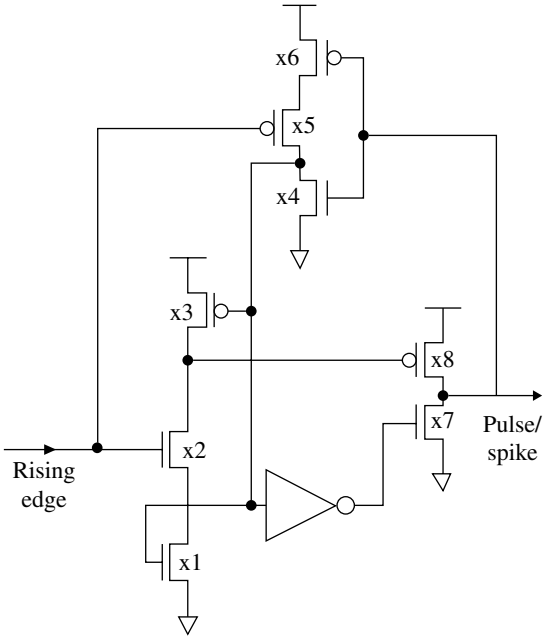


FIGURE 14.21 Circuit diagram of axon hillock model [14]. Reproduced with permission of IEEE.

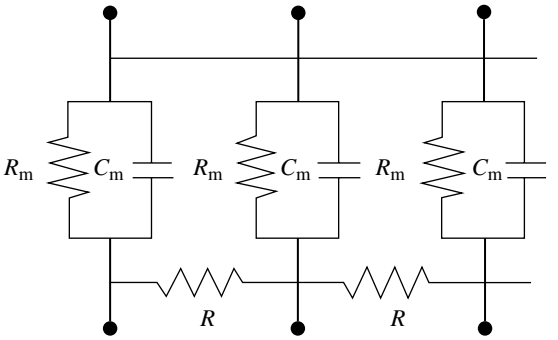


FIGURE 14.22 Circuit diagram of axon model.

The simulation result is shown in Figure 14.23. The R_m and C_m are assigned to simulate the demyelinated and normal axon transmission, and all compartments have identical values. The axon output is plotted in a transient simulation. The demyelinated output shows obvious signal loss.

14.20 CONCLUSIONS AND FUTURE TRENDS

We have designed a circuit with a CNT synaptic layer and CMOS neurons that mimics hyperactivity in the cortex and thalamus in an OCD affected brain, and shows how a "accelerator–brake mechanism" controls the level of activity in the brain. This circuit simulation can be used to probe further by strengthening and weakening synaptic strengths to demonstrate more complex behaviors. Varying many more mechanisms in the neuromorphic circuits is possible, including firing threshold, receptor availability, and neurotransmitter availability and reuptake rate, to demonstrate additional physiological behavior.

The human brain is a prediction machine. It detects regularities in environmental stimuli and stops responding to them. If the input is not what is expected, humans can typically comprehend this difference between reality and expectation. For example, if someone listens to his favorite music, he can predict the next note or beat. Also when two people are in conversation, there is synchrony between them that helps each to predict the next word of the other [3]. Failure in predictions or wrong responses can cause hallucination in schizophrenia [1]. Research shows that alteration in neural circuitry can cause disinhibition in pyramidal neurons and hence can cause hallucinations.

We have designed a circuit that mimics a small circuit in a healthy brain and schizophrenic brain and its responses. This circuit uses CNT transistors for the synaptic layer and TSMC 45 nm MOSFET technology for neurons. There are emerging applications for this hybrid architecture in bio-nanotechnology. This hybrid architecture combines inherent advantages of CMOS and CNTs.

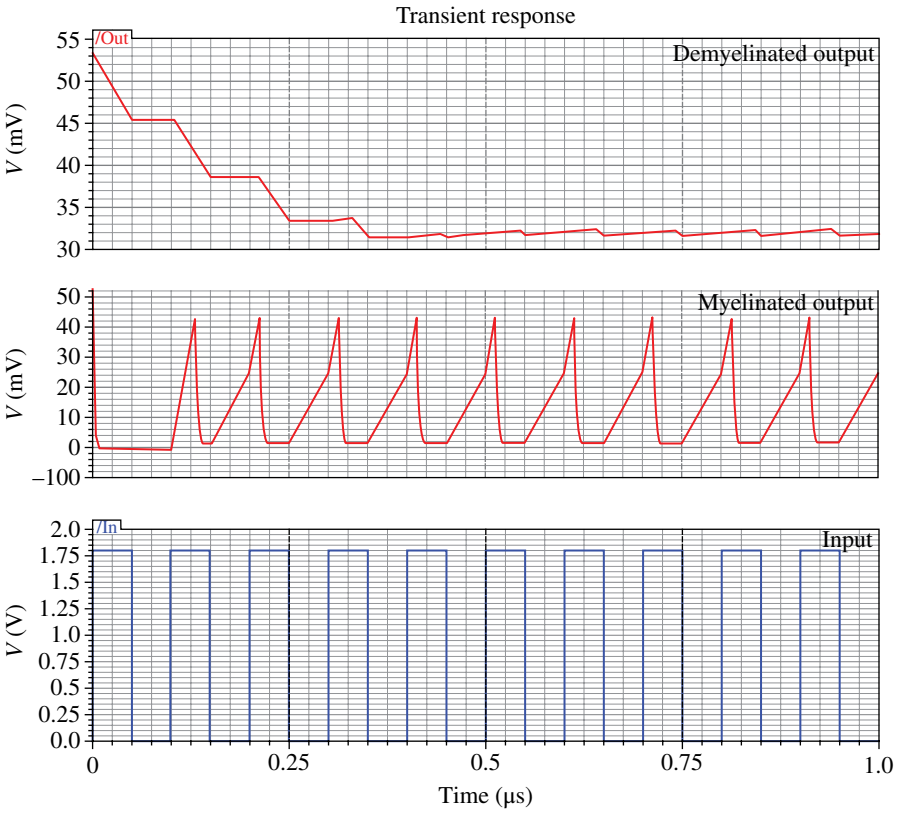


FIGURE 14.23 A system block diagram of the neuron with axon model.

It uses CNT transistors as biological sensors and CMOS transistors as complex electronic control and processing unit [11].

We are capable of simulating a portion of fundamental processing in PD by neuromorphic engineering and achieving a similar output signal to biological neural pulses. In the future, a complete circuit model will help us understand the complexities of PD integrating multiple effects, theory, or even treatment. Furthermore, a bioelectronic neuromorphic circuit with partial normal function may be an implantable treatment for incurable neurology disease, such as PD.

The research described here is a beginning, a tentative step toward modeling human behavior, both typical and atypical. While the networks here are vastly simplified over human neural networks, some basic behavior concerning prediction and mistaken perception has been demonstrated here. Researchers can run simulations with our circuits, changing synapse strengths, neurotransmitter availability, and other parameters, to see how the behavior changes, and correlate it with observations from biological measurements.

REFERENCES

[1] J. Joshi, Chih-Chieh Hsu, A. C. Parker, and P. Deshmukh, "A carbon nanotube cortical neuron with excitatory and inhibitory dendritic computations," in *2009 IEEE/NIH Life Science Systems and Applications Workshop*, Bethesda, MD, IEEE, Piscataway, NJ, pp. 133–136, April 9–10, 2009.

[2] C. Mead, Analog VLSI and Neural Systems. Reading, MA: Addison-Wesley Publishing Company, 1989.

[3] A. K. Friesz and A. C. Parker, "A biomimetic carbon nanotube synapse," Proceedings of the BMES Annual Fall Meeting, 2007, pp. 1–5.

[4] A. C. Parker, "The BioRC project website." Available at: http://ceng.usc.edu/parker/BioRC_research.html (accessed on January 12, 2016).

[5] J. Joshi, J. Zhang, C. Wang, Chih-Chieh Hsu, A. Parker, C. Zhou, and U. Ravishankar, "A biomimetic fabricated carbon nanotube synapse for prosthetic applications," in *Life Science Systems and Applications Workshop (LiSSA), 2011 IEEE/NIH*, Bethesda, MD, IEEE, Piscataway, NJ, pp. 139–142, April 7–8, 2011.

[6] J. Joshi, A. C. Parker, and Chih-Chieh Hsu, "A carbon nanotube cortical neuron with spike-timing-dependent plasticity," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, Minneapolis, MN, IEEE, Piscataway, NJ, pp. 1651–1654, September 3–6, 2009.

[7] J. Joshi, A. C. Parker, and T. Celikel, "Neuromorphic network implementation of the somatosensory cortex," in *IEEE EMBC Conference on Neural Engineering*, San Diego, CA, IEEE, Piscataway, NJ, pp. 907–910, November 6–8, 2013.

[8] J. Joshi, Plasticity in CMOS neuromorphic circuits. Ph.D. thesis, University of Southern California, Los Angeles, CA, 2013.

[9] Chih-Chieh Hsu, Dendritic computation and plasticity in neuromorphic circuits. Ph.D. thesis, University of Southern California, Los Angeles, CA, 2015.

[10] Chih-Chieh Hsu, A. C. Parker, and J. Joshi, "Dendritic computations, dendritic spiking and dendritic plasticity in nanoelectronic neurons," in *53rd Midwest Symposium on Circuits and Systems, 2010. MWSCAS 2010*, Seattle, WA, IEEE, Piscataway, NJ, pp. 89–92, August 1–4, 2010.

[11] K.-C. Tseng, A. C. Parker, and J. Joshi, "A directionally-selective neuromorphic circuit based on reciprocal synapses in starburst amacrine cells," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Boston, MA, IEEE, Piscataway, NJ, pp. 5674–5677, August 30–September 3, 2011.

[12] K.-C. Tseng, Neuromorphic sensing circuits in a silicon retina. Ph.D. thesis, University of Southern California, Los Angeles, CA, 2012.

[13] M. Mahvash, Emulating variability in the behavior of artificial central neurons. Ph.D. thesis, University of Southern California, Los Angeles, CA, 2012.

[14] M. Mahvash and A. C. Parker, "Synaptic variability in a cortical neuromorphic circuit," IEEE Transactions on Neural Networks and Learning Systems, 24, 397–409, 2013.

[15] Y. Irizarry-Valle and A. C. Parker, "Astrocyte on neuronal phase synchrony in CMOS," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, Melbourne, VIC, IEEE, Piscataway, NJ, pp. 261–264, June 1–5, 2014.

[16] J. Joshi, A. Parker, and K.-C. Tseng, "*In-silico* glial microdomain to invoke excitability in cortical neural networks," in *2011 IEEE International Symposium on Circuits and Systems*, Rio de Janeiro, IEEE, Piscataway, NJ, pp. 681–684, May 15–18, 2011.

[17] Y. Irizarry-Valle, A. C. Parker, and J. Joshi, "A CMOS neuromorphic approach to emulate neuro-astrocyte interactions," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Dallas, TX, IEEE, Piscataway, NJ, pp. 1–7, August 4–9, 2013.

[18] A. C. Parker, J. Joshi, Chih-Chieh Hsu, and N. A. D. Singh, "A carbon nanotube implementation of temporal and spatial dendritic computations," in *51st IEEE Midwest Symposium on Circuits and Systems*, Knoxville, TN, IEEE, Piscataway, NJ, pp. 818–821, August 10–13, 2008.

[19] C. L. Chen, K. Kim, Q. Truong, A. Shen, Z. Li, and Y. Chen, "A spiking neuron circuit based on a carbon nanotube transistor," Nanotechnology, 23, 275202, 2012.

[20] J. T. Ting and G. Feng, "Glutamatergic synaptic dysfunction and obsessive-compulsive disorder," Current Chemical Genomics, 2, 62–75, 2008.

[21] S. E. Ahmari, T. Spellman, N. L. Douglass, M. A. Kheirbek, H. B. Simpson, K. Deisseroth, J. A. Gordon, and R. Hen, "Repeated cortico-striatal stimulation generates persistent OCD-like behavior," Science, 340, 1234–1239, 2013.

[22] J. Davenport, "Diagnostics and Statistical Manual of Mental Health Disorders (IV)," 4th ed. Washington, DC: American Psychiatric Association, 1994.

[23] S. B. Math and Y. C. Janardhan Reddy, "Issues in the pharmacological treatment of obsessive-compulsive disorder," International Journal of Clinical Practice, 61, 1188–1197, 2007.

[24] R. L. Albin, A. B. Young, and J. B. Penney, "The functional anatomy of basal ganglia disorders," Trends in Neurosciences, 12, 366–375, 1989.

[25] A. R. Crossman, "Primate models of dyskinesia: The experimental approach to the study of basal ganglia-related involuntary movement disorders," Neuroscience, 21, 1–40, 1987.

[26] R. G. Robertson, S. M. Farmery, M. A. Sambrook, and A. R. Crossman, "Dyskinesia in the primate following injection of an excitatory amino acid antagonist into the medial segment of the globus pallidus," Brain Research, 476, 317–322, 1989.

[27] E. Kandel, J. Schwartz, and T. Jessel, Principles of Neural Science, vol. 4. McGraw-Hill, New York, 2004.

[28] R. D. Fields, "Brain wiring," *Psychology Today*, 2016. [Online]. Available: https://www.psychologytoday.com/blog/the-new-brain/201106/brain-wiring [accessed: Februsry 15, 2016].

[29] J. Deng and H. S. P. Wong, "A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application—Part I: Model of the intrinsic channel region," IEEE Transactions on Electron Devices, 54, 3186–3194, 2007.

[30] D. Akinwande, S. Yasuda, B. Paul, S. Fujita, G. Close, and H. S. P. Wong, "Monolithic integration of CMOS VLSI and CNT for hybrid nanotechnology applications," in *38th European Solid-State Device Research Conference, 2008. ESSDERC 2008*, Edinburgh, IEEE, Piscataway, NJ, pp. 91–94, September 15–19, 2008.

[31] C. Chen, D. Xu, E. S. Kong, and Y. Zhang, "Multichannel carbon-nanotube FETs and complementary logic gates with nanowelded contacts," IEEE Electron Device Letters, 27, 852–855, 2006.

[32] S. Dikker, L. J. Silbert, U. Hasson, and J. D. Zevin, "On the same wavelength: Predictable language enhances speaker–listener brain-to-brain synchrony in posterior superior temporal gyrus," The Journal of Neuroscience, 34, 6267–6272, 2014.

[33] M. Bar, "The proactive brain: Using analogies and associations to generate predictions," Trends in Cognitive Sciences, 11, 280–289, 2007.

[34] G. Horga, K. C. Schatz, A. Abi-Dargham, and B. S. Peterson, "Deficits in predictive coding underlie hallucinations in schizophrenia," The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 34, 8072–8082, 2014.

[35] F. M. Benes, "Emerging principles of altered neural circuitry in schizophrenia," Brain Research. Brain Research Reviews, 31, 251–269, 2000.

[36] M. G. Vučković, Q. Li, B. Fisher, A. Nacca, R. M. Leahy, J. P. Walsh, J. Mukherjee, C. Williams, M. W. Jakowec, and G. M. Petzinger, "Exercise elevates dopamine D2 receptor in a mouse model of Parkinson's disease: In vivo imaging with [f]fallypride," Movement Disorders, 25, 2777–2784, 2010.

[37] K. T. C. Ardith, G. Remington, and E. Frohman, "Multiple sclerosis," Medical Clinics of North America, 93, 451–476, 2008.

[38] J. Joshi, A. Parker, and Chih-Chieh Hsu, "A carbon nanotube spiking cortical neuron with tunable refractory period and spiking duration," in *IEEE 2010 Latin American Symposium on Circuits and Systems (LASCAS 2010)*, Iguaçu Falls, Brazil, IEEE, Piscataway, NJ, February 24–26, 2010.

15

LINKING MEDICAL NANOROBOTS TO PERVASIVE COMPUTING

SYLVAIN MARTEL
Department of Computer Engineering, Polytechnique Montréal, Montreal, QC, Canada

15.1 INTRODUCTION

Medical microrobots [1] and medical nanorobotics [2] are relatively recent areas of intense researches. The main difference between microrobotics and nanorobotics is the fact that the field of medical nanorobotics exploits nanotechnology with the principles of robotic to provide new tools and methods in the medical field. As such, a nanorobot is defined here as a microscale robotic entity or an agent that has nanometer-scale components in order to integrate functionalities that are only possible by exploiting phenomena that only occur at the nanoscale. Although it could potentially result into significant impacts on future approaches for early detections and diagnostics, initial research efforts have been done toward the enhancement of treatment outcomes in cancer therapy [3]. The potential use of medical nanorobotics in cancer therapy is significant when one looks at the statistics for the United States alone. Indeed, for the year 2013 alone, 1,660,290 new cancer cases have been estimated resulting in approximately 580,350 deaths. The latter represents almost 1,600 people per day or 1 every 54 s. With an estimated today's world population of 7.119 billion individuals with US population representing only 4.45%, one can easily realize the importance of finding new or improved treatment modalities.

Knowing that most cancers are spatially localized in the form of solid tumors, medical nanorobotics can contribute in improving cancer treatments by providing additional complementary functionalities to target specific areas better, which would

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

lead to optimized therapeutic effects while reducing systemic circulation of passively circulating agents that most often lead to an increase of the toxicity for the patient. To perform effectively, these nanorobotic agents must presently be supported and controlled by upgraded and/or special computerized interventional platforms. Because of technological constraints for implementing traditional communication methods at the scales of such nanorobotic agents, other approaches are being used to link such agents to larger computerized platforms that in turn are linked to computer networks, hence providing the fundamental communication links for integrating and connecting medical nanorobotic agents to pervasive computing.

15.2 COMPLEMENTARY FUNCTIONALITIES

Medical nanorobotics can contribute in improving cancer treatments by providing three main additional complementary functionalities to existing therapeutic agents: navigation, actuation, and sensory capabilities. Actual tumor targeting relies mainly on passive or active targeting [4]. Passive targeting relies on the accumulation of therapeutic agents in pathological sites with compromised vasculature [5]. It relies on systemic circulation that results in a lower therapeutic index, which leads to an increase in toxicity with a lower therapeutic efficacy. Active targeting is based on the attachment of specific ligands to the surface of pharmaceutical carriers to recognize and bind to pathological cells. Again, this approach also relies on systemic circulation.

The systemic vascular network is made of close to 100,000 km of blood vessels in each human adult. Although the total length of such a network corresponding roughly to two and a half times the circumference of planet Earth at the equator provides access to any parts in the body, it also provides pathways to healthy organs and tissues that are negatively affected by highly efficient, hence highly toxic, pharmaceutical agents.

In medical nanorobotics, the concept of direct targeting [2] is introduced instead. For direct targeting, pharmaceutical carriers or agents are being navigated directly from the injection site to the target area in order to avoid or at least reduce systemic circulation of highly toxic molecules. This principle of direct targeting is represented schematically in Figure 15.1.

Figure 15.1 describes the difference using a tree-like network between systemic-based deliveries and direct targeting using navigable therapeutic agents. As depicted in the figure, systemic circulation affects the whole network whereas for direct targeting, only the targeted region is affected. Such principle of direct targeting needs compatible navigable agents capable of transporting therapeutic molecules. This is discussed in more detail later in this chapter.

The second complementary functionality provided by medical nanorobotics is actuation. Actuation in the form of directional propelling force is used not only to navigate such agents in the vascular network but also to pass the diffusion limit of larger therapeutic molecules in tumoral tissues. Indeed, there is a lack of flow in the

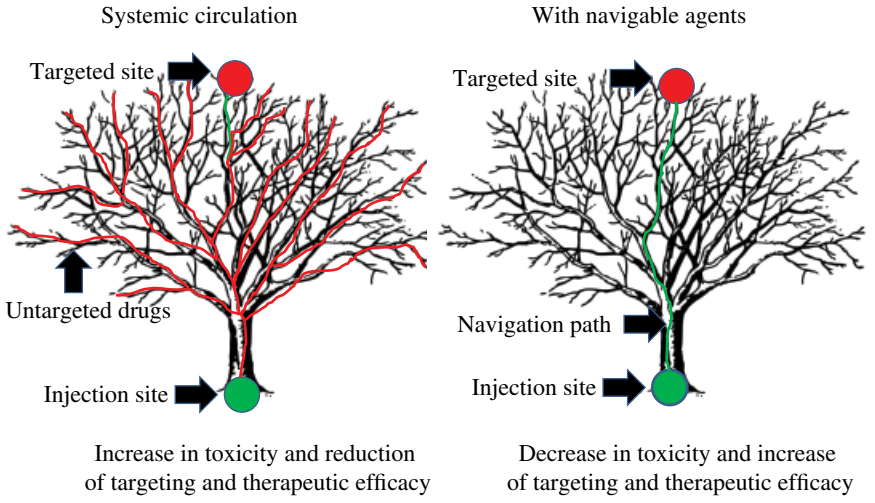


FIGURE 15.1 Systemic-based targeting versus direct targeting.

tumor due to what is known as the tumor interstitial fluid pressure (TIFP). This TIFP prevents therapeutic molecules to penetrate deeper in the tumor, hence limiting their therapeutic effect [6]. By providing some form of propulsion, medical nanorobotic agents have the capability to counteract such TIFP and to penetrate deep into solid tumor through the interstitial spaces.

The third complementary functionality that can be provided by nanorobotic agents is in the form of sensors. Tumors for instance are not homogeneous and may contain hypoxic regions where the cancer cells are more active. Hypoxic regions are characterized by a low oxygen concentration due to the consumption of oxygen by active and duplicating cancer cells. Targeting such regions would enhance further the therapeutic efficacy of the dose being delivered [7]. Using embedded oxygen sensors to follow decreasing oxygen gradients from the blood vessels to the hypoxic regions inside tumors is one example where medical nanorobotic agents can play an important role in cancer therapy.

15.3 MAIN SPECIFICATIONS FOR SUCH NANOROBOTIC AGENTS (NANOROBOTS)

Medical nanorobotics must make use of navigable agents. Especially in the microvasculature where no medical imaging modality exists to visualize such vessels in order to guide them along the proper vascular routes, such agents must become more sophisticated, demonstrating some form of intelligence as typically encountered in traditional robots. But the implementation of such "intelligent" nanorobotic agents requires the integration of fundamental specifications that can be extremely challenging and most likely impossible to

implement at such a scale with present technology for several of them. These specifications are as follows:

1. The physiological environment determines the maximum overall size of the nanorobotic agents, and for targeting a tumor they must be capable of following the same route as the oxygen and nutrients that feed the tumor. This means that such agents must access the tumor through the angiogenic network since the TIFP prevents direct injections to the tumor itself. More specifically, these nanorobotic agents must be able to transit through all potential physiological routes leading to the tumor. These routes include the chaotic irregular angiogenic network made of vessels with a diameter as low as 4 μm, the interstitial spaces, and the intercellular openings of typically less than 2 μm that are responsible for much of the blood vessel leakiness in solid tumors [8]. As such, the maximum diameter of the nanorobotic agents should be approximately 2 μm.
2. Due to technological limitations to induce an appropriate displacement force such as a magnetic force on such small agents considering the distance from the actuation source located outside the patient, self-propelled nanorobotic agents would typically be required at such a scale. Hence, each agent must have a propulsion system that performs effectively in low Reynolds number regime on the order of 10^{-4} as encountered in such physiological microenvironments.
3. Such embedded propulsion system must also be actuated by one or more motors capable of providing sufficient force to transit effectively through the potential physiological routes, leading to the tumor and deep inside the tumor interstitial space well beyond the diffusion limit of conventional drug molecules.
4. Since no medical imaging modalities with sufficient spatial resolution exist to gather images of the microvascular networks and other potential routes to reach the tumor, an appropriate level of autonomy must be embedded in each nanorobotic agent since they cannot rely on an external source to help them in navigating around physiological obstacles. As such, each nanorobotic agent must have some form of obstacle detector connected to some sort of onboard computer being linked to a steering system.
5. Although each agent could be able to navigate autonomously through physiological obstacles, they must have a form of onboard receiver (Rx) navigation system to allow a computerized external imaging platform to indicate the general direction of the target being a tumor in this particular application. Because of the overall size of a tumor when being treated, imaging and determining the location of the tumoral region using existing clinical imaging modalities are possible. This embedded Rx navigation system must be connected to the onboard computer and steering system in order to influence the directional displacement in an efficient and appropriate manner to achieve the required targeting and therapeutic efficacy.

6. Once in the interstitial space of a solid tumor, such nanorobotic agents must deliver their therapeutic payloads at locations where the therapeutic outcomes would be maximized. Indeed, a tumor is generally not homogeneous but contains regions where cancer cells are more active and duplicate faster. Because such cellular activity consumes oxygen, they are characterized by hypoxic regions of low oxygen concentrations of approximately 0.5%. Therefore, to maximize the therapeutic outcomes, each nanorobotic agent must be able to localize and deliver their therapeutic payloads within these hypoxic regions. Such autonomy being embedded in each agent is required since such hypoxic regions cannot at the present time be localized using external systems including existing medical imaging modalities. This in turn requires that each nanorobotic agent would have an oxygen (O_2) sensor connected to the onboard computer in order to control the steering system or the directional displacement of the agent, following a decreasing oxygen gradient from the angiogenesis network where the oxygen concentration is maximal to such hypoxic regions inside the tumoral volume.
7. Each nanorobotic agent must also have an oxygen threshold detector being set at approximately 0.5% oxygen in order to stop the displacement of the nanorobotic agents in the hypoxic regions until the drug molecules are being released. Without such oxygen threshold detector, the nanorobotic agents would bypass the hypoxic regions and potentially reach the necrotic regions (dead cells) where the therapeutics would result into no real contribution toward the treatment outcomes. Such oxygen threshold detector (0.5% O_2 detector) must be connected to the onboard computer to stop the propelling system or to control the propelling system for back-and-forth displacements in order to maintain each nanorobotic agent within the hypoxic regions.
8. Each nanorobotic agent must have an embedded power source to maintain all these onboard systems and functionalities operational for a sufficient length of time required to complete the target intervention.
9. Each nanorobotic agent and hence their embedded components must also be biocompatible and meet both cytotoxicity and immune system response requirements for potential use in humans.
10. Finally, each nanorobotic agent must be able to carry sufficient therapeutic payloads encased in several nanometer-scale containers to avoid direct exposure and excessive leakage of the toxic drug molecules prior to reach the target regions that need treatment. The drug-loaded containers should be attached to the surface of each microscale nanorobotic agent. As such, each agent must have sufficient surface to attach a sufficient quantity of such drug-loaded containers, suggesting a diameter just below 2 μm for round-shaped agents.

Although the aforementioned specifications would allow such nanorobotic agents to perform their tasks effectively, in a practical point of view, additional specifications would be required to allow such technologies to reach the clinics. For instance,

because of their small sizes, only a very small amount of therapeutics can be carried by each nanorobotic agent. Therefore, a relatively large number of such agents would be required. The quantity of nanorobotic agents required to deliver a sufficient dose to achieve an adequate therapeutic effect could be in the tens to hundreds of millions nanorobotic agents. Therefore, to offer affordable treatments based on such nanorobotic agents, typical controlled assembly methods are not well suited for such applications. Instead, such agents considering the multiple functionalities being embedded should ideally be self-replicating. Furthermore, the attachment of the drug-loaded containers should rely on a self-assembly process.

15.4 MEDICAL NANOROBOTIC AGENTS—AN EXAMPLE

The displacement force for such nanorobotic agents acting in low Reynolds hydrodynamic conditions such as in the tumor environments can be done by induction from an external source or by embedding a self-propelling system. For the former, magnetic torque is presently known to be the best option in the microvasculature since it requires much less power than for magnetic pulling methods that do not scale well to perform an effective navigation beyond the arteries and in larger arterioles. Magnetic induction can be made on microstructures known as artificial swimmers that mimic the propulsion systems of small organisms such as the flagella in bacteria that proved to be highly efficient in such low Reynolds hydrodynamic conditions. Even if artificial bacteria flagella (ABF) [9] for instance could potentially be build in mass production at a relatively low cost with biocompatible materials, the level of embedded autonomy required to achieve the highest therapeutic index by targeting specific zones such as the hypoxic regions would be beyond technological feasibility at such a scale. The same is true for self-propelled agents in the form of small rockets or catalytic propulsion microsystems [10] that besides the need for a biocompatible fuel would require external supports that cannot presently fulfill all the requirements listed in the preceding section.

Since developing an artificial medical nanorobotic agent that fulfills all the specifications listed in the preceding section is well beyond actual technological feasibility, one logical alternative has been to consider a microorganism characterized with the right specifications including biocompatibility that would act as a biological nanorobotic agent and which could be interfaced and controlled with a computerized platform located outside the patient. In other words, an organic nanorobotic agent could be considered instead of an inorganic implementation, which would be well beyond technological feasibility. As such, the MC-1 magnetotactic bacterium (MTB) [11] has been initially considered to act as such organic medical nanorobotic agent [12, 13] instead of nonrobotic-assisted conventional bacterial therapies [14]. The MC-1 cell, as depicted in Figure 15.2, is round shaped with a diameter ranging from approximately 1 to 2 μm being as discussed earlier, appropriate to transit through potential physiological routes and inside solid tumors. Each MC-1 cell has two flagella bundles shown at the lower side of the cell in Figure 15.2 that provide the required thrust propulsive force [15] in low Reynolds hydrodynamic conditions to

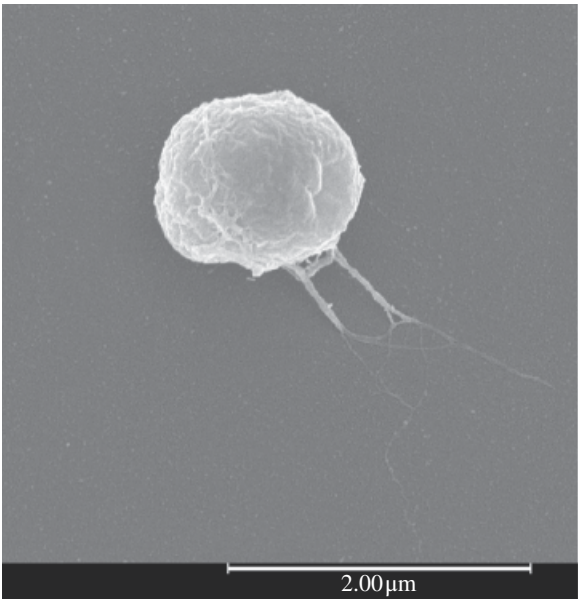


FIGURE 15.2 Microscopy image of the MC-1 magnetotactic bacterium.

reach the tumor and to penetrate deep inside the tumor interstitial space well beyond the diffusion limit of typical drug molecules or therapeutic agents relying on passive or active targeting alone. These flagella are actuated by molecular motors where each motor has a rotor capable of rotating 360° inside a stator. Such design, which is similar to larger modern-engineered electrical motors, allows anticlockwise rotations to move forward and clockwise rotations to move backward.

Although many microorganisms could fulfill many of the specifications listed in the previous section, some others such as the Rx navigation system and the oxygen sensor with a corresponding threshold detector are other specifications that are not always supported adequately. To provide the required functionalities to target the hypoxic regions, the migration behavior of the MC-1 cells is exploited. Indeed, in their natural environment, the MC-1 cells rely on magneto-aerotaxis migration behavior [16] to survive. In other words, a chain of membrane-based single magnetic domain iron oxide nanocrystals known as magnetosomes [17–19] in each cell are used to direct the bacteria toward regions of lower oxygen concentrations, which are located deeper in water or sediments. As such, they swim along the lines of geomagnetic field where the latter induces a directional torque on the chain of magnetosomes acting like a miniature magnetic compass needle. During their migration, they sense the oxygen concentration until they find the appropriate oxygen concentrations for their survival. Because the MC-1 cells are microaerophilic, they seek low oxygen concentrations and will form microaerophilic bands at approximately 0.5% oxygen, which, as mentioned earlier, corresponds approximately to the oxygen concentrations found in hypoxic regions inside solid tumors.

For clinical applications and to emulate the required control that would be performed on equivalent but futuristic medical artificial nanorobots, an associated computerized platform known as the magnetotaxis platform [20] is used to exploit their magneto-aerotaxis migration behavior to target such hypoxic regions in solid tumors. Unlike as it is the case in their natural environment, for targeting the hypoxic regions, the magnetotaxis component is first exploited independently of the aerotaxis component since the latter cannot initially contribute to the targeting efficacy and to benefice from the known location of the tumor mass when these bacteria operate outside the tumoral volume. As such, a three-dimensional volume referred to here as the aggregation zone (AZ) is typically placed at the tumor location. Such AZ is somewhat similar to an artificial magnetic pole. When outside the AZ, the strength of the magnetic field pointing toward the AZ and generated by pairs of electromagnetic coils assembled in the magnetotaxis platform surrounding the patient is sufficient to induce a directional torque on the chain of magnetosomes in each drug-loaded MC-1 cell. (Although other forms of payloads are possible, presently approximately 70 nanoliposomes at 170-nm molecular containers known as liposomes are attached per MC-1 cell using a self-assembly biochemical method and are used to encase the drug molecules that are released once at the target location [21].) But once inside the AZ, the field strength becomes too weak to direct the motion of the bacteria. Any bacteria exiting the AZ will be forced to reenter the AZ due to the induced directional torque. This process results into an accumulation or aggregation of the MC-1 cells in the AZ. Once in the AZ, since magnetotaxis is no more effective, aerotaxis takes effect where the MC-1 bacteria follow the decreasing oxygen gradients to form aggregates in the hypoxic regions of the tumors.

Once in the tumor volume, the patient can be transported to a medical-imaging modality. In this particular case, a clinical magnetic resonance imaging (MRI) scanner would be the most appropriate modality due to the fact that the iron oxide magnetic nanoparticles inside each bacterium act as MRI contrast agents and hence make the location of the bacteria visible if the concentration of bacteria is sufficiently high, which should typically be the case considering that a large population of bacteria is required to deliver sufficient drug molecules in order to achieve an adequate therapeutic effect. Imaging such population of drug-loaded bacteria provides valuable information to the medical specialists such as assessing the targeting efficacy while providing an estimate of the amount of therapeutics that have been delivered in the treated regions. This is even more critical when using more efficient drug molecules that show more toxicity since it could, in the cases of less successful targeting processes, prevent further injections of highly toxic agents that could impact healthy tissues and organs, which could be fatal for the patients.

15.5 NANOROBOTIC COMMUNICATION LINKS ALLOWING PERVASIVE COMPUTING

The communication links from such bacterial nanorobotic agents and a computer network are fundamentally achieved through two basic mechanisms depending upon the direction of the communication link, while the type of communication is

presently limited but sufficient to allow the execution of relatively high sophisticated tasks. In both directions, the communication link is wireless and is based on magnetic fields that are exploited in different manners. From the computer network to the MC-1 nanorobotic agents, the type of communication is typically constrained to indicate the direction of motion to the nanorobotic agents. The associated command cannot be sent to a single nanorobotic agent or subgroups of agents but rather to all nanorobotic agents present in the dedicated interventional space. With a very high number of agents that can easily reach for instance tens of millions of individuals in a single injection for the delivery of drug molecules in cancer therapy, independent addressing to each of these agents in such high population would yield excessive communication overheads, which would be impractical in this context. The transmission system in this particular case is the magnetotaxis platform that exploits the magnetotactic displacement behavior of the MC-1 cells through the generations of magnetic fields aimed at directing such population of nanorobotic agents toward the desired location. As mentioned previously, this is achieved by generating an artificial pole at the target location such as the tumor volume to entail these bacteria acting as medical nanorobots to migrate toward the target region by exploiting the magnetotactic component of their natural magneto-aerotactic migration behavior.

In the reverse direction from the nanorobotic agents to the computer network, the communication receiver would typically take the form of a clinical MRI scanner while the source of the receiving signal will mainly be based on the distortion of the B_0 field (the high magnitude homogeneous (uniform) magnetic field in the tunnel of the MRI scanner) created by magnetosomes inside the MC-1 cells. The feasibility of such communication link depends not only on the characteristics of the magnetosomes, the configuration of the chain of magnetosomes, and on the density and the overall size of the MC-1 cell population, but also on the MRI sequence being used and the sensitivity of the clinical MRI scanner that is acting as a communication receiver in this particular context.

15.6 TYPES OF INFORMATION

The types of information that could be made available from such organic nanorobotic agents in the context of pervasive computing can take various forms. The first obvious type of information is related to the therapeutic index. Here, the therapeutic index defines the ratio of the drug-loaded MC-1 nanorobotic agents that reached the target region requiring treatment versus the ratio of drug-loaded MC-1 agents that did not reach the target and which could contribute to increase the level of toxicity for the patient through systemic circulation of highly toxic drug molecules affecting healthy organs and tissues. Since the use of such nanorobotic agents is likely to enhance significantly the therapeutic index compared to systematically injected procedures such as chemotherapy, more efficient and hence more toxic drug molecules could be envisioned. On the other hand, a mistake during such interventions such as setting the AZ outside the region that needs treatments could lead to fatal consequences to the patient, especially when using highly toxic drug molecules. Hence, being able to monitor or assess the targeting efficacy to determine if other injections can be pursed

based on the therapeutic index of the previous injections, without causing any risks for the patient, becomes mandatory or at least highly suitable. As such, the location and the intensity of the MRI signal being gathered and processed using special algorithms to determine the quantity of drug-loaded bacterial nanorobotic agents at the target site can be used to determine the therapeutic index from the known quantity of drug being carried per bacterial carrier and the number of bacterial carriers that has been injected.

Other types of information that can be retrieved from such bacterial carriers are related to diagnostics and imaging. For instance, hypoxic regions in solid tumors cannot presently be targeted or visualized with modern medical technologies. The exploitation of the microaerophilic behavior of the MC-1 cells leads to results similar to the ones that could be achieved by futuristic inorganic nanorobots equipped with oxygen sensors and capable of following a decreasing oxygen gradient prior to maintain their position within the low oxygen concentration encountered in hypoxic regions in solid tumors. The MRI signals being retrieved from the artifacts created in the B_0 field by the magnetosomes inside the MC-1 cells emulating such artificial futuristic nanorobots can also be used to gather information of the hypoxic zones, which can be useful in assessing the progression and the characteristics of the tumor, which in turn can help determine the best treatment modality.

Diagnostic applications can be expanded by attaching molecules with the right specificity. The detection sensitivity of the imaging modality can be enhanced through the use of markers attached to the surface of the bacterial agents. For example, iron oxide nanoparticles can be attached to the surface of MC-1 cells to amplify the MRI artifacts from a smaller population of bacteria. Other hybrid implementations include but are not limited to nanocomponents such as gold nanoparticles that can be attached to the surface of the MC-1 cells to enhance treatments based on radiotherapy since such treatment modality is much less effective in hypoxic regions due to the lack of oxygen. Multitreatment modalities such as combining target radiotherapy and chemotherapy by attaching two different therapeutic agents to each bacterial cell or by injecting two populations of MC-1 cells, each with a single type of therapeutic agent, are also feasible. The latter can also be combined with diagnostics to implement theranostic nanorobotic agents exploiting the two-way communication links between the bacterial nanorobotic agents and the computer networks. These examples only provide a feel about the many possibilities that can be presently supported using existing organic and hybrid nanorobotic agents.

15.7 MEDICAL NANOROBOTIC AGENTS FOR MONITORING AND EARLY DETECTION

Although the field of medical nanorobotics is progressing at a fast pace, nanorobotic agents still rely on massive infrastructures and platforms to conduct medical interventions. The need for large and bulky systems such as a clinical MRI scanner and a magnetotaxis platform to provide supports and the communication links between the nanorobotic agents and the computerized platforms are presently an

obstacle to the wide use of pervasive computing in this particular field. Indeed, systems providing the computing power necessary to support the tasks performed by such nanorobotic agents can be easily installed on-site such as in the interventional room without the space and power constraints of portable computing devices that must rely on pervasive computing to compensate for a lack of sufficient on-site computing power.

But one of the future medical applications where pervasive computing could be a substantial addition to the use of nanorobotic agents would be in the early detection of diseases. Indeed, it is well recognized that the overall cost of health care could be decreased substantially if the early detection of diseases could be achieved at relatively low costs. Such concept has already been applied for cars by implementing sensors at various strategic locations within each vehicle. The same concept could be applied for humans. In this case, two main approaches could be envisioned. The first one would rely on a set of various sensors located at various locations within the body. Such sensors would not move but remain at the same locations permanently. Excluding biocompatibility and acceptance issues, this approach would work if the signal-to-noise ratio (SNR), assuming that the sensibility of such sensors would be adequate, would be sufficiently high. But to achieve earlier detections, the sensitivity of each sensor must be increased. Higher sensitivity will then translate to earlier detections with the possibility of treating the patients at the earliest time, hence providing the maximum chance to achieve the best treatment outcomes. Another issue is where to place such highly sensitive sensors. In some cases, signs of a particular disease are spread throughout the body where in other cases, specific areas considered as high risks are identified and monitored. But fulfilling these two conditions will not provide the fundamental approach to realize early detection capabilities for all types or at least a much wider range of potential diseases that could occur anywhere within the body.

These facts suggest a second approach making use of sensors capable of displacement within the body and as such, medical nanorobotics could become a suitable option. Indeed, using the large vascular network as a transport medium, every part of the body could be reached. This transport medium could also be expanded through other physiological spaces including but not limited to the interstitial spaces that all would be reachable through the blood vessels and in particular the capillary networks that are constituted of narrowed blood vessels. These vessels and other physiological sites impose serious constraints on the overall maximum size of these movable sensors or sensory-based medical nanorobotic agents and other potential sensors. Such a small size typically translates into weak signals that affect the overall sensibility of the approach aimed at performing early detections. For instance, it is well known that reducing the surface area of an electrode increases the electrical impedance, hence resulting in a reduction of its sensitivity. To increase its ability to detect weaker physiological changes through an increase in its sensitivity for a given conductive material on the surface of the electrode, the surface of the electrode must be increased. Similarly, the signal being transmitted in early detection applications can be amplified to facilitate the detection by an external system by the use of many sensor-based medical nanorobotic agents within the same physiological site.

As such, three fundamental strategies could be envisioned, being referred to here as systemic, regional, and dispatched approaches. In the systemic approach, such sensor-based nanorobotic agents would drift through the systemic blood circulation and accumulate at a specific site when encountering specific conditions or specificities corresponding to the early warning sign of a specific disease. Actuation force could be exploited in some cases to reach regions where diffusion alone would be inefficient. To reduce the searching effort and the overall time required to perform the search while avoiding systemic circulation, a regional approach could be envisioned. The latter would assume that we would have an a priori knowledge of the regions where the risks of the occurrence of a particular disease would be much higher. This is often the case in cancer, for instance. Indeed, it is difficult for cancer cells to survive outside their region of origin, so they must accumulate and grow in another physiological environment having similar characteristics. For instance, colon cancer has a tendency to metastasize to the liver while stomach cancer metastasizes to the ovary in women. Therefore, it could be envisioned that both the colon and the liver would be prioritized in the case of possible occurrences for someone who survived colon cancer for instance. But the most powerful approach in medical nanorobotics aimed at monitoring the whole body for the early detection of particular diseases is most likely the dispatched approach. In such a mode of operation, information is gathered and used to dispatch the nanorobotic agents to a specific site at the earliest possible time. Such information could be gathered from various sources including the nanorobotic agents themselves.

15.8 MEDICAL NANOROBOTICS AND PERVASIVE COMPUTING—MAIN CONDITIONS THAT MUST BE MET FOR ITS FEASIBILITY

As mentioned earlier, constant monitoring to achieve early detection of diseases could potentially benefice from pervasive computing since such type of applications would bypass the need for frequent scheduled examinations that do not monitor the status of the human body on a permanent basis and where a relatively long period of time can separate two successive appointments. But there are many conditions that must be met before such a concept connecting medical nanorobotics with pervasive computing with the aim of monitoring and providing early detections within the human body becomes a reality. Presently, the MC-1 bacteria acting as such nanorobotic agents are nonpathogenic, and as such these microorganisms only survive up to approximately 30 min in the physiological conditions encountered in the human body. Although such time period is sufficient in most cases for medical applications such as drug delivery to tumors, this limited survival time is a major obstacle to perform the long-term monitoring required to support the early detection of diseases.

The use of organic nanorobots would require that the agents would be able to duplicate themselves or perform some sort of self-assembly within the body. Such ability is typically supported by pathogens and as such, some sort of control of the population growth would be required. Such control could be provided by the changes in physiological conditions that would be linked to a particular disease. For instance,

the tumor environment differs from the rest of the body. Therefore, such a difference could be exploited to trigger and enhance or facilitate the rate at which the population of the nanorobotic agents would grow. Such growing population would contribute to enhance the sensory signals and facilitate the early detection of the disease or the beginning of a tumor in this particular example. Going further, therapeutic effects could be combined with the detection of the disease. For example, magneto-aerotactic bacteria could be engineered to grow in the low oxygen concentration of hypoxic regions in solid tumors. The increase in population of bacteria would consume the oxygen required for the cancer cells to grow, which in turn when such oxygen would be limited would reduce the bacterial population. Such approach would avoid the need for each bacterial nanorobotic agent to carry therapeutic molecules suitable to each type of tumor or disease and which could not be synthesized and attached to new bacterial cells when in the human body. But one issue to resolve would be to find a mechanism by which a portion of the bacterial nanorobotic agents would survive and operate in all other physiological sites with higher-oxygen concentrations. On the other hand, synthetic, artificial, or inorganic medical nanorobotic agents would most likely not have the capability of reproducing themselves or to perform self-assembly of other nanorobotic agents, at least in a relatively near future. Hybrid implementations might represent a potential alternative in many cases. But it is too early to predict such a scenario with certitude.

Assuming that such nanorobotic agents can be implemented and operated within the human body, the next issues would be concerned with communication links. Although the initial research effort in the field as mentioned for the MC-1 bacterial nanorobotic agents suggests potential methods based on magnetism to communicate in an environment exploiting pervasive computing, miniaturization of the bulky platforms in the form of low-powered portable devices will be essential. As such, new sophisticated algorithms may help in achieving such a goal, which in turn may rely on communication with other computing devices to achieve the required computing power to execute such advanced algorithms. The same may hold true for performing tasks such as advanced diagnostics.

15.9 CONCLUSION

Medical nanorobotics is indeed progressing at a fast pace, and preliminary results suggest its potential in enhancing treatments, imaging, and diagnostics. Such nanorobotic agents can combine diagnostics and therapies to form theranostic nanorobotic agents. The use of nanorobotics for medical applications is a very powerful concept when one realizes that many new features and characteristics can enhance and improve the outcomes in medical diagnostics and treatments. The contribution of medical nanorobotics results in new complementary functionalities such as navigation, actuation, and sensory-based displacements. Already, relatively sophisticated nanorobotic agents capable of performing target therapeutic tasks have been implemented in the form of organic nanorobots. Artificial nanorobots still have technological constraints that prevent the implementation of some of the functionalities

embedded in some organic nanorobots. But as technology evolves, artificial or hybrid versions of such medical nanorobots may become more suitable for particular medical tasks. Nonetheless, the level of computation embedded in each of these nanorobotic agents will remain limited. Although biochemistry and other specialized molecules, for instance, can reduce the need to implement many computerized functions, extending the level of computing power available would most likely be beneficial for the proper execution of many medically related tasks. Some forms of swarm behavior through interactions between the nanorobotic agents may offer a suitable option. In other cases, relying on external computing devices through pervasive computing may also be suitable in particular tasks. Although some forms of communication links already implemented and tested between nanorobotic agents and an external computer have suggested potential avenues to implement such communication networks between future nanorobots and computing devices in a pervasive computing environment, many issues need to be resolved before such paradigm becomes effective.

REFERENCES

[1] Nelson, B.J., Kaliakatsos, I.K., and Abbott, J.J. 2010. Microrobots for minimally invasive medicine. Annual Review of Biomedical Engineering, 12, 55–85.

[2] Martel, S. 2014. Magnetic therapeutic delivery using navigable agents. Therapeutic Delivery, 5(2), 189–204.

[3] Martel, S. 2013. Navigation control of micro-agents in the vascular network: Challenges and strategies for endovascular magnetic navigation control of microscale drug delivery carriers. IEEE Control Systems, 33(6), 119–134.

[4] Torchilin, V.P. 2010. Passive and active drug targeting: Drug delivery to tumors as an example. Handbook of Experimental Pharmacology, 197, 3–53.

[5] Greish, K. 2010. Enhanced permeability and retention (EPR) effect for anticancer nanomedicine drug targeting. Methods in Molecular Biology, 624, 25–37.

[6] Milosevic, M.F., Fyles, A.W., Wong, R., Pintilie, M., Kavanagh, M.C., et al. 1998. Interstitial fluid pressure in cervical carcinoma: Within tumor heterogeneity, and relation to oxygen tension. Cancer, 82, 2418–2426.

[7] Brown, J.M. and Wilson, W.R. 2004. Exploiting tumour hypoxia in cancer treatment. Nature Reviews Cancer, 4, 437–446.

[8] McDonald, D.M. and Baluk, P. 2002. Significance of blood vessel leakiness in cancer. Cancer Research, 62, 5381–5385.

[9] Zhang, L., Abbott, J.J., Dong, L., Peyer, K.E., Kratochvil, B.E., Zhang, H., Bergeles, C., and Nelson, B.J. 2009. Characterizing the swimming properties of artificial bacterial flagella. Nano Letters, 9(10), 3663–3667.

[10] Soler, L., Martinez-Cisneros, C., Swiersy, A., Sánchez, S., and Schmidt, O.G. 2013. Thermal activation of catalytic microjets in blood samples using microfluidic chips. Lab on a Chip, 13, 4299–4303.

[11] Blakemore, R.P. 1975. Magnetotactic bacteria. Science, 190, 377–379.

[12] Martel, S., Mohammadi, M., Felfoul, O., Lu, Z., and Pouponneau, P. 2009. Flagellated magnetotactic bacteria as controlled MRI-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature. International Journal of Robotics Research, 28(4), 571–582.

[13] Martel, S. 2013. Bacterial microsystems and microrobots. Biomedical Microdevices, 14(6), 1033–1045.

[14] Patyar, S., Joshi, R., Prasad Byrav, D.S., Prakash, A., Medhi, B., and Das, B.K. 2010. Bacteria in cancer therapy: A novel experimental strategy. Journal of Biomedical Science, 17(21), 1–9.

[15] Hughes, M.P. and Morgan, H. 1999. Measurement of bacterial flagellar thrust by negative dielectrophoresis. Biotechnology Progress, 15(2), 245–249.

[16] Frankel, R.B., Bazylinski, D.A., Johnson, M.S., and Taylor, B.L. 1997. Magneto-aerotaxis in marine, coccoid bacteria. Biophysical Journal 73, 994–1000.

[17] Bazylinski, D.A., Frankel, R.B., and Jannasch, H.W. 1988. Anaerobic magnetite production by a marine, magnetotactic bacterium. Nature, 334, 518–519.

[18] Faivre, D. and Schüler, D. 2008. Magnetotactic bacteria and magnetosomes. Chemical Reviews, 108, 4875–4898.

[19] Schuler, D. 1999. Formation of magnetosomes in magnetotactic bacteria. Journal of Molecular Microbiology and Biotechnology, 1, 79–86.

[20] de Lanauze, D., Felfoul, O., Turcot, J.-P., Mohammadi, M., and Martel, S. 2014. Three-dimensional remote aggregation and steering of magnetotactic bacteria microrobots for drug delivery applications. International Journal of Robotics Research, 33(3), 359–374.

[21] Taherkhani, S., Mohammadi, M., Daoud, J., Martel, S., and Tabrizian, M. 2014. Covalent binding of nanoliposomes to the surface of magnetotactic bacteria acting as self-propelled target delivery agents. ACS Nano, 8(5), 5049–5060.

16

NANOMEDICINE'S TRANSVERSALITY: SOME IMPLICATIONS OF THE NANOMEDICAL PARADIGM

José J. López and Mathieu Noury*
Sociological and Anthropological Studies, University of Ottawa, Ottawa, ON, Canada

16.1 INTRODUCTION

This chapter is concerned with developing an analysis of the implications of the nanomedical paradigm. Although the nanomedical paradigm promised revolutionary changes, actual developments appear to have fallen short of the more radical transformations that had been foreseen. In this chapter, we begin by reviewing some of the recent literature that has assessed the extent to which nanomedicine has lived up to its promise. The general consensus is that despite encouraging signals, the field has yet to mature. One of attempting to develop an analysis of the implications of an emerging field such nanomedicine is that its inchoate and yet-to-be-determined nature makes extrapolation of its impact highly unfeasible. We draw on scholarship from science studies to argue that despite their compelling nature, the future visions of nanotechnology and nanomedicine, advanced by early promoters, rested on a narrow technological determinist understanding of technological development and innovation. By ignoring the highly contextual nature of technological development, these initial raw visions of nanomedicine's future have failed to grasp how the technology on the ground would develop. Equally, we draw on this same literature to show that the preliminary anticipatory visions, which fuelled the development of the

* Names appear in alphabetical order.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications,
First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

fields of nanotechnology and nanomedicine, have to be understood less as actual predictions of future development, and more as attempts to pragmatically intervene in the present to create the technical, epistemological, funding, and political conditions for the development of the field. Judged from this optic, that is, the reorganization enabling the emergence and consolidation of the field, there can be no doubt that nanomedicine has been enormously successful as we argue later.

In drawing attention to the perils of extrapolation and prediction we do not, however, argue that attempts to grasp the implications of nanomedicine should be abandoned. Instead, we argue that more nuanced and cautious efforts to monitor the developmental pathways of the different technologies associated with the nanomedicine need to be developed. We thus propose our analysis not as a prediction of future developments, but as a way of sensitizing researchers and policy makers to some of the choices that are already being made, wittingly or not, as the diverse technologies develop. In particular, we draw attention to the manner in which the nanomedicine paradigm is perhaps already contributing to changing our conception of illness and health and what it means to be an embodied being. We organize our discussion around the notion of the transversality of nanomedicine and its molecular basis. We show how across the three most important areas of contemporary biomedicine—predictive, personalized, and regenerative medicine—nanomedicine has crosscutting effects. In addition we highlight that despite its indisputable novelty, the nanomedicine paradigm in some areas, particularly predictive and personalized medicine, builds on and intensifies already-existing tendencies within biomedicine. We conclude by recapitulating a paradox associated with each of the three areas and argue that these paradoxes provide useful sites for monitoring and reflecting on the development of nanomedicine.

16.2 NANOMEDICINE'S PROMISES

Globally supported by public policy and investment [1] nanomedicine, defined as the use of nanotechnologies in medicine, is presented as a medical revolution that promises to radically change the practice of health care "from diagnostic to therapeutic, and everything in between" [2]. In a recent appraisal, however, Duncan and Gaspar argue that the novelty and radical break that nanomedicine is expected to engender is frequently overplayed: "Some seem unaware of the historical background and in their enthusiasm overexaggerate (hype) giving the impression that their proposed new technologies are already reality today" [3]. Moreover, it is worth recalling, for instance, that 15 and 10 years, respectively, have transpired since the passage of the US National Nanotechnology Initiative as well as the National Institute of Health released its *Roadmap* of the *Nanomedicine Initiative*. In the latter, it was claimed "'this cutting-edge area of research will begin yielding medical benefits as early as ten years from now', specifically mentioning 'molecular machines' that will change the landscape of medicine" (cited in Ref. [4], p. 1). Few commentators would deny the fertility of nanomedicine as a field (see Ref. [3] for an important and incisive overview) when seen through the prism of research

activity and funding. However, judged in terms of clinical success and paradigm shifts, as Venkatraman notes, "the results appear more modest" [4] (p. 1). Even more optimistic commentators are forced to admit that the field remains in a state of "relative adolescence," and that "a large portion of nanomedicine applications identified are still in the research and development state" [5] (p. 11). Indeed as Duncan and Gaspar highlight:

> Too many publications still begin with a phrase "widely used for biomedical applications" whereas in most cases the system has not even entered clinical trial or been tested in vivo, and even some first generation technologies that are in clinical development are not yet products approved for routine human use. Investigators need to understand and articulate the differences between a lab experiment and a medicine if society is to reap the benefits [3] (p. 2118).

A possible explanation for this inability to distinguish between "lab experiment" and "medicine" can be found in Juliano's [6] analysis of the obstacles encountered by the field's multidisciplinary orientation that is worth quoting at length:

> A major issue is the difference in mindset or intellectual culture between nanotechnologists and biomedical scientists. The chemists, physicists and engineers who design and produce the nanodevices used in nanomedicine are trained to deal with physical systems and machines…Some machines may be very intricate…[but]…Once built, a machine does not change the way it functions and its behaviour is usually predictable by mathematical equations. Not so with biological systems which can adapt to changing circumstances and whose behaviour is hard to predict. These characteristics of biological systems present conceptual challenges to physical scientists. The biomedical scientists who study living systems are comfortable with their complexities, but they are often uncomfortable with the precise and quantitative tools used to describe and manipulate physical systems. Since nanomedicine is inherently a multidisciplinary field, there is inevitably a mismatch of intellectual cultures between the technologists and the biologists [6] (pp. 100–101).

More broadly, Juliano also concludes that "leaders in the field of nanomedicine should exercise restraint and not overpromise concerning the extent to which nanotechnology will change healthcare, or the rapidity with which those changes will be introduced" [6] (p. 102).

In the closing remarks of the European Technology Platform (ETP) *NanoMedicine Vision Paper*, published in 2005, it is claimed that "In the future, a nanoparticle or a set of nano particles may be designed to search for, find and destroy a single diseased cell, taking us even closer to realising the ultimate goal of disease prevention" [7] (p. 35). A strategic positioning of nanomedicine vis-à-vis science fiction prefaces these comments: "Future techniques in medical diagnosis and treatment have often been the subject of science fiction literature and cinema. What was once the stuff of science is now closer to becoming a reality" [7] (p. 35). Although the demarcation between nanotechnology and science fiction has been frequently invoked to distinguish legitimate from junk science [8], it still remains the case that science fiction can be

conjured more positively to evoke a future transformed by the potential associated with current technological capabilities:

> science fiction is often deployed to produce a sense of wonder about science, to capture the unbelievable, the amazing. Charting the drama of the seeming inevitable movement from science fiction to science, it reveals a fascination with new scientific developments and the wonders they make visible, wonders previously perceptible only in sf [science fiction]. Scientists themselves evoke the idea of science fiction as a way of capturing the incredible speed of technoscientific change [9] (p. 271).

However in its 2013 White Paper, the ETP in Nanomedicine (ETPN) accepts that while European funding in the area of nanomedicine has contributed to the development of novel nanomedicines, few of these products have been taken to market [7], thus concurring with the assessments of the field presented before. The document certainly continues to highlight the significance of the nanoscale as the Archimedean point from which new therapeutic and diagnostic applications will arise and even "revolutionise the early diagnosis of diseases and their therapy" [7] (p. 8). The tone of the white paper, however, takes a decisively nonscience fiction register. It draws attention to the more mundane tasks of incubating a culture of "open innovation" by developing relationships between the various stakeholders, most significantly academic researchers and industry: "The best option to successfully implement the open innovation model in Europe consists in establishing a supply chain providing nanomedicine products compatible with industrial processes and strategies" [7, 10] (p. 4).

This of course does not mean that the hope or the expectation that nanomedicine will transform medicine has dissipated. Indeed as Anderson and her colleagues show, scientists and policy makers in the broad field of nanoscience, including nanomedicine, are generally optimistic [11]. Similarly, despite the lack of maturity in the field, Etheridge and colleagues still assert, "continued development … should lead to the truly revolutionary advances in medicine" [3, 5] (p. 11). Moreover, Duncan and Gaspar, for all their unease with the fields' penchant for hype, still conclude "harnessing the undoubted potential [of nanomedicine] will make a major contribution to improved healthcare in the 21st century" [3] (p. 2128). Nonetheless, these assessments do point to the fact that technical potential narrowly conceived, that is, the myriad ways scientists are able to intervene at the nanoscale length, is not itself a guarantee of the ability to scale up and translate novel discoveries into diagnosis, therapy, or cure. A corollary statement, which we develop in more detail later, is that much caution should be exercised when using such "raw" technical potential as the springboard for extrapolating the future effects of current technological and technical possibilities.

In his review of the future of nanomedicine, Juliano notes, "the nanomedicine literature is replete with publications that describe clever, elegant innovations in nanotechnology accompanied by projections of therapeutic or diagnostic utility. However many of these technologies fail to move forward toward practical utilization" [6] (p. 103). The prospect that innovative nanotechnologically enabled medical

discoveries might run aground in inadequate policy, and social environments, however, should have hardly caught policy experts by surprise. From the very beginning, nanotechnology, and its subfields such as nanomedicine, was the product of significant policy intervention that sought to actively and decisively remodel the research and funding environment to sustain the emergence and growth of research in nanotechnology [12]. First and most important was the development of new funding programs in the United States, Europe, Canada, and Japan, often fuelled by international rivalry [12–14]. This enabled the kind of focused multidisciplinary research that has come to characterize nanotechnology. Following directly from this was the redesignating of "established fields of research as 'nano-science', thus bringing them within the purview of the new nanotech initiatives" [13] (p. 147). Such reorganizing meant that well-established work that had been undertaken before the consolidation of nanotechnology proper under the rubric of "applied chemistry," "surface physics," "macromolecular physics," and "supramolecular chemistry" was re-categorized as nanoscience or nanotechnology [13]. In addition, it also reincorporated disciplines such as physics, engineering, and materials science into significant funding streams from which they had been marginalized in the biotechnology bubbles of the 1980s and 1990s [14] (p. 193). Consequently, commentators have claimed that much of the alleged novelty of nanotechnology was "largely the result of incremental advances in already well-established branches of applied science, such as material science and colloid technology" [15] (p. 1). Indeed, recent bibliometric research suggests that despite evidence of their multidisciplinary nature, in the fields of nanotechnology and nanoscience the bulk of research is concentrated in the "macrodisciplines of material science and chemistry" [16]. Although some scientists and policy makers describe the emergence of nanotechnology and science as a mere rebranding of existing science [11] (p. 140), there can be little doubt that the various nanotechnology strategies and initiatives in the United States, and abroad, have largely been successful. They have generated genuinely new funding structures, research institutions, and research projects that probably would not have developed absent the nanotechnology umbrella [12, 13].

16.3 ANALYSING IMPLICATIONS OF THE NANOMEDICINE PARADIGM

In the area of nanomedicine, progress is discernible. This includes innovations such as new *in vivo* and *in vitro* diagnostic tools (quantum dots, golden nanoparticles, implants, lab-on-a-chip, etc.), advanced drug delivery systems, innovative procedures combining diagnostics and therapy (theranostics), and even new tissue engineering methods. That these successes do not, however, correspond with the sweeping transformations predicted by the more enthusiastic promoters should not be taken as a sign of the failure of nanomedicine. To understand why this is, it is necessary to turn to the social sciences. Over the past decades, a considerable body of theoretical and empirical work has developed in the field of science studies (see Ref. [17] for an overview). A significant area of research has been the study of the broader social,

political, cultural, and economic processes in which technology and technological innovation are embedded. Amongst the field's findings, two are pertinent for understanding why the more thorough radical promises remain, and are perhaps, unrealizable in the terms in which they were formulated. Equally, the literature also suggests avenues for more cautious modes of anticipating the development of the emerging technologies associated with nanomedicine. The first concerns the explanatory weakness of the linear model of technological innovation that was implicit in the extrapolation of nanoscience and nanomedicine's future at the launch of the varied nano-initiatives. This is known as the technological determinist model, namely, where it is the inner logic or potential of the technology itself that explains the impact it will have on society. In other words, once you identify what a particular technology can do in a laboratory, you can then predict its broader impact on society, the economy, manufacturing, medicine, etc. Instead a more complex mode of analysis for technological development needs to be adopted. This is because

> Technology does not develop according to an inner technical logic but is instead a social product, patterned by the conditions of its creation and use. Every stage in the generation and implementation of new technologies involves a set of choices between different technical options. Alongside narrowly "technical" considerations, a range of "social" factors affect which options are selected – thus influencing the content of technologies, and their social implications [18] (p. 866).

Consequently, once technological development is decoupled from the idea of "a predetermined logic or a single determination," innovation can better be reconceptualized as a "garden of forking paths" in which "different routes are available, potentially leading to different technological outcomes" [18] (p. 866). In other words, initial conditions, be they technological or social, can never provide sufficient information to predict technological outcomes or social impacts in the future. The development of technologies has to be reflexively monitored in real time [19] in order to ascertain how technological and social conditions recursively select and reselect new developmental pathways. The highly contextual nature of the innovation process explains the difficulty of accurately extrapolating technological developments into the future. This results from the "ramifications" [20] that are instantiated as choices, both conscious and unconscious, are made when technologies take shape on the ground.

Important as the critique of the technological determinist model is for understanding why initial predictions would most likely be inaccurate, it does not, however, account for the reasons why the visions of the future associated with the emergence of nanoscience and nanomedicine have taken the form of proclamations of radical transformations. One thing is to linearly extrapolate the consequences of a technology into the future; another is that such inferences take the form of radically transformed and highly stylized future societies, where everybody's interest is optimized. In the case of nanotechnology, these types of future visions "occur beyond the usual venues of popular science and science fiction; they are also present in more technical material as well as governmental and institutional policy documents" [21] (p. 2) [8, 22].

This brings us to the second pertinent finding from the science studies scholarship. Theoretically informed empirical analysis of the emergence of new technologies or scientific research programs reveals that the function of promises of future societal transformations, which accompany such cases, is not to predict the future. Indeed, time and time again such predictions have been proven erroneous [23, 24]. Rather than provide a realistic vision of the future,

> initial promises are set high in order to attract attention from (financial) sponsors, to stimulate agenda setting processes (both technical and political) and to build "protected spaces". Promises thus play a role in the social processes that are part of technological development…Some future speculations do not strive for truth or accuracy, but are meant to influence specific social processes in technological developments [23] (p. 881).

While expectations in the form of promised, or hoped for, impact, have always been a key dimension of technological development processes, contemporary visions are characterized by their "hyperbolic expectations of future promise and potential" [24] (p. 286). This is due to a variety of factors, including the strategic as opposed to serendipitous management of scientific and technological innovation, the alignment of innovation with economic growth and national or regional competitiveness, and the increased complexity and cost of contemporary innovation systems [24]. Moreover, as our discussion of the ETPN already makes clear, while

> it used to be common to make a distinction between science and technology, the former concerned with "pure" scientific discovery whilst the latter with translating the discoveries of science into socially useful and in many cases saleable artefacts. Increasingly science scholars argue that the distinction must be abandoned. Insofar as the majority of scientific endeavours are currently legitimated by their more immediate social contribution, we have entered the reign of technoscience. Science may still remain a juggernaut in contemporary society, but the pursuit of knowledge for knowledge's sake is no longer sufficient to power it. Pure science, if pure ever it was, it is now incontestably hybrid [22] (p. 12).

This has an impact not only in terms of the broader configuration of funding schemes and research structures, but it also drills down to the level of actual laboratory practices, linking "forward-looking statements" and "local articulations of scientific research" [25]. The reasons why an understanding of the implications of nanomedicine cannot take as its point of departure either the inflated promissory notes issued by its most ardent promoters, or be understood as an exercise of mere extrapolation of an inner technical logic can be summarized as follows. In the case of the former, grand visions of the future are not acts of clairvoyance; instead, they are "future-orientated abstractions" that perform pragmatic work in the present: "they guide activities, provide structure and legitimation, attract interest and foster investment. They give definition to roles, clarify duties, offer some shared shape of what to expect and how to prepare for opportunities" [24] (p. 286). In the case of the latter, the shaping of technological potential and innovation via interaction of social, cultural,

political, and economic variables means that narrowly conceived technological extrapolations will most certainly fail to predict the shifting trajectories of the technologies as they develop on the ground.

In pointing out the weaknesses associated with these two modes of projecting nanomedicine into the future, our objective is not to be unduly pessimistic with respect to our ability to govern these technologies, or worst yet to suggest that such attempts are futile due to the radical uncertainty of the future. Instead, an awareness of the pitfalls associated with too clearly seeing the beneficent effects of nanomedicine's future—(typically, a world much like ours but only better!)—should contribute to a more reflexive and grounded understanding of the potentials associated with nanomedicine as it develops in real time. In this sense, the work undertaken by David H Guston and his team at the Center for Nanotechnology in Society at Arizona State University is of particular relevance [26]. The goal of the diverse projects at the center is not to engage in speculative *NanoEthics* [27], prediction, or foresight. It is to analyze "plausible futures with emphasis on such methods as scenario development that provide a more diverse and normative vision compared with other methods that seek to identify a single, most likely future" [26] (p. 226). This potentially sensitizes researchers and policy makers to the fact that there are opportunities to shape and mould nanotechnology as it develops at its multitudinous sites, perhaps contributing to "bending the long arc of technoscience more toward humane ends" [26] (p. 234).

It is in this context that we inscribe our analysis of the implications of the nanomedicine paradigm in the remaining sections of this chapter. Unlike the work undertaken by Guston and his colleagues, our analysis does not derive from formal scenario development workshops or social media platforms exploring the future of emerging products [26, 28]. Instead, it draws on interviews across a number of nanomedical research sites [29]. The intent of our analysis is not to predict the future of nanomedicine but to raise questions regarding the social choices that are consciously or unconsciously being incorporated as nanomedicine's future is being imagined today and as specific nanomedical technologies mature. Guston and his team mobilize the term "anticipation" to make clear that their mode of analysis should not be confused with "expectation, prediction, or foresight"; instead drilling into the etymology of the word Guston argues that anticipation "is instead related to 'capable' and 'capacity', from the Latin *capere*, meaning to take into possession. With the prefix 'ante-', meaning 'before' 'with regard to position, order or time…anticipation is thus more about practicing, rehearsing, or exercising a capacity in a logically, spatially, or temporally prior way than it is about divining a future" [26] (p. 226). Anticipation then is the development of a versatile reflexive capacity through which stakeholders explore and closely scrutinize assumptions and knowledges as a way of grasping tenable developmental trajectories of concrete technologies. Similarly, exploring the etymology of "implication" also allows us to express what our analysis is meant to convey. "Implication" from the Latin *implicare* and *implicatio* means to "intertwine, entangle and involve" [30]. Thus in using the term "implications," we are attempting to point to the manner in which both the nanomedicine paradigm as it is emerging, and specific technological innovations as they are becoming concretized are intertwined with particular understandings of what health, disease, and the body are or

should be. These particular entanglements of technologies, norms, ideals, and representations of health and illness provide us with cognitive tools with which to explore plausible developmental trajectories, and perhaps even with insights that might enable us to better steer the development of the varied trajectories within the nanomedicine paradigm.

Whereas many of the researchers that we have cited earlier project nanomedicine's transformative power into the future when the different nanotechnologically enabled products are incorporated into everyday medical practice, we would like to suggest that detectable in the technologies as they are currently being developed is the potential for important transformations in our conception of and eventually the lived reality of health, disease, and the body. Once again, we stress these are not predictions. They are attempts to draw attention to some of the choices being made frequently unwittingly as nanomedicine matures. By raising them, we hope to contribute to broadening the range of issues and variables that are taken into consideration as different stakeholders participate in the collective governance of nanomedicine.

Analyses of policy documents, interviews with nanomedical scientist, and scientific publications led one of the authors to propose analyzing nanomedicine in terms of its *transversality* [29, 31]. The transversal nature of nanomedicine reflects its ambition to lay the foundations for and to become the catalyst of a new biomedical era where medicine as such would be understood as *nano*medicine. The projected outcome of this paradigm shift is to include under the *nano* umbrella the key components of contemporary biomedicine: predictive, personalized, and regenerative medicine. In other words, nanomedicine is envisioned as a crosscutting model of health care that weaves together these three key domains in contemporary medicine, setting the stage for a more coherent and integrated future model of medicine organized around the nanoscale, a major paradigm shift in patients' care and medical intervention. This future vision contains an upgraded version of contemporary medicine, optimized for detecting the molecular origins of disease and efficiently curing and repairing. While aspects of these claims are plausible, we are not concerned with, and in any case cannot adjudicate, the veracity of these claims, the likelihood of their becoming a reality in the future, or the form that they will actually take. We are interested in the implications, as defined before. In other words, for us the significance of the nanomedicine futures, which are currently circulating, does not rest on their accuracy in predicting the future. Instead, we are interested in exploring the extent to which "scientific expectations of the 'future of nanomedicine' may take on a material dimension, becoming variously incorporated into diagnostic procedures, treatment options, and new biomedical technologies" [25] (p. 46) as they become entangled with new norms of health and illness. As a first step in this process, we begin by discussing what we understand to underpin nanomedicine's transversality: its molecular focus. Whereas many scholars point to the radical break that the nanomedicine paradigm represents with respect to existing biomedical conceptions and practices, we argue that some fundamental features of nanomedicine, as it is developing, can also be grasped in terms of its continuity with and the deepening of modern medicine's molecularization.

16.4 THE MOLECULAR UNDERPINNINGS OF NANOMEDICINE'S TRANSVERSALITY

It is now a shared theoretical and practical assumption among the biomedical community that health and illness are best grasped at the molecular scale insofar as the normal and the pathological are both ontologically understood as arising from molecular mechanisms and processes [32]. This assumption is rooted in what the sociologist Nikolas Rose calls "a profound 'molecularization' of styles of biomedical thought, judgment, and interventions" [33] (p. 11). The origin of this shift is to be found in biology's efforts to visualize life at the submicroscopic scale and to adopt a new language appropriate to such a scale, through the use of metaphorical terms such of genetic "code," "information," and "program" [34–36]. As a result of its dominance in biology, biomedicine came to adopt this molecular gaze, which has significantly transformed the modes of medical reasoning and their representation of disease. In other words, molecularization is the starting point of an epistemological and ontological shift in biomedicine that takes its roots in biology's "molecular turn," and has, consequently, profoundly changed both medical practice's object and practical rationality. It is because of this change, argues Rose, that contemporary biomedicine now "envisages life at the molecular level, as a set of intelligible vital mechanisms among molecular entities that can be identified, isolated, manipulated, mobilized, recombined, in new practices of intervention, which are no longer constrained by the apparent normativity of a natural vital order" [33] (p. 6).

Nanomedicine is today the most tangible and advanced achievement of the molecularization of biomedicine. Its transversal vision of health care proposes a new paradigmatic conception of biomedical practice built on the model of a molecular medicine that integrates the domains of predictive, personalized, and regenerative medicine. In this sense, the cornerstone of the nanomedical paradigm, and the condition of its transversality, is to be found at the size scale of the nano-interventions: the nanometer or the subcellular scale. This is why nanomedicine is regarded as working at "the ideal scale" for medicine [37]. As Kewal K. Jain noted, "physiological and pathological processes at cell level occur on the nanoscale" [38] (p. 90). Indeed, the molecular focus is what underpins, and thus justifies the breadth of the nanomedical spectrum and explains its transversal or crosscutting nature. Widely used in the discourses articulating nanomedicine's promises, the following quote from the Nobel Laureate physicist Richard E. Smalley is a good example of the significance of the molecular vision: "Human health has always been determined on the nanometer scale; this is where the structure and properties of machines of life work in every one of the cells in every living thing. The practical impact of nanoscience on human health will be huge" (cited in Ref. [39] (p. 1)). This same vision expressed more prosaically provides the lynchpin that underpins the nanomedical endeavor:

> It is interesting to note that biologic systems are inherently composed of nanoscale building blocks. The width of a DNA molecule is approximately 2.5 nm. The dimensions of most proteins are in the range of 1.0 to 15.0 or 20.0 nm, and the width of cell membranes is in the range of 6–10 nm. […] basic life processes take place on the nanoscale [40] (p. B25).

The origin of nanomedicine's "huge" potential is tightly correlated to an understanding of human health that is conceptually dependent on a molecular understanding of the body as a biological system. It is because "basic life processes take place at the nanoscale" that nanomedicine is expected to transform the whole field of biomedicine and healthcare innovation. The prospect of devising new materials, devices, and architectures that are able to more precisely identify and control the molecular processes underpinning pathology constitutes the bedrock of the nanomedical paradigm. A similar vision of the prospect of molecular control underwrote the human genome project, and indeed until fairly recently most scholars writing about the molecularization of health had genetics, and latterly genomics in mind. However, it is worth drawing attention to the change in the meaning of "molecularization" as it passes from the field of biotechnology to that of nanomedicine in which the former is increasingly being subsumed. Amongst the metaphors used to convey the potential associated with the human genome project were a "plan," "blueprint," "encyclopedia," "code" or "program of life" [41]. In all the cases, the imagined futures made possible by these metaphors encountered the unsurpassable frontier between the organic and the inorganic. However, the lay of the land is different for nanotechnology and nanomedicine. This is because the nanometer not only constitutes the scale "where the structure and the machines of life work," but also the spectrum at which the world might be shaped one-atom-at-a-time, making nanotechnology to return to Smalley once again, "the builder's final frontier" [42]. Thus whereas the molecular vision implicit in genetics and genomics mapped on to "life," in nanotechnology and consequently nanomedicine it is coextensive with nature itself, making life "programmable matter" [43]. As a result, the body is understood as "matter" in two different yet interrelated senses:

> First, the body is taken as natural matter, as the biological, living body that interacts constantly with nonbiological, nonliving objects – this would constitute the body as the object for nanomedical intervention. But the body is also taken as technical matter, in a reductionist move that states that, at bottom, both living and non living, biological and nonbiological entities are essentially matter. This latter view opens the material world – and the body – to the affordances of instrumentality at the atomic-molecular level [43] (p. 122).

As we will show later, nanomedicine not only intensifies and accelerates the molecularization that was already present in genetics and genomics with respect to predictive and personalized medicine. It also has the potential in the case of regenerative medicine to significantly change how we think about our embodied existence.

16.5 NANOMEDICINE AS PREDICTIVE MEDICINE

The ultimate goal of nanodiagnostics is to identify disease at the earliest possible stage through the detection of molecular biomarkers [7], which may be roughly defined as indicators whose presence suggests a potential pathological development [44]. As the medical anthropologist Margaret Lock noted, many researchers consider

molecular biomarkers to be the actual essence of a future disease [45]. The generalized expectation within the field is that the ability to identify molecular biomarkers will produce a reliable predictive knowledge of the future development of specific pathologies, enabling earlier and more targeted interventions, leading to disease prevention. Though not immediately apparent, this probabilistic form of medical knowledge goes beyond preventive medicine's classical epidemiological reasoning. It focuses on presymptomatic signs suggesting that, despite appearances to the contrary, a disease process exists in a state of virtuality: "nanoscale diagnostics are expected to identify in the becoming, giving the opportunity to intervene specifically prior to a symptomatically detected onset disease" [46] (p. 2). Nanodiagnostic reasoning is particularly concerned with identifying predictive indicators of disease through the detection of molecular risk factors associated to particular diseases. Indeed, through emerging applications, such as nanochips, quantum dots, or nanobiosensors [47], nanomedicine aims to produce new diagnostic tools where, in the absence of an actual pathology, the present biological condition might still yield predictive signs of future abnormalities. In other words, it seeks to "be ahead" of the disease [29, 31] (p. 109).

Consequently, nanodiagnostics does not merely look for molecular markers in order to reveal or confirm an existing disease, but to predict disease development. Biomarkers are considered to be an integral part of the pathological process; they indicate the possible "presence" of pathology before the latter is manifest. When microarray analysis detects a specific protein expression associated with breast cancer, it predicts a preexisting cancer that has not yet developed. Because disease is associated with potentiality, molecular analysis confronts us with a paradoxical situation where the successful discovery of evidence of disease rests on its actual absence [48]. The objective existence of the disease is defined according to a probabilistic projection based on molecular risk factors analysis. However, some argue that such highly predictive reasoning remains extremely limited because predicting the transition from genotype to phenotype, and the severity and the age of disease outbreak, is almost impossible for multifactorial diseases that are most common [49].

In an interview for *MIT Technology Review*, biologist Leroy Hood noted that nanodevices are in reality geared toward "interrogating health rather than disease" [50], meaning that nanomedicine focuses less on fighting disease than on detecting factors that could threaten health at some indeterminate time in the future. In this sense, nanodiagnostics expresses a paradigm shift both in the temporality and the location of medical intervention, that is, before the appearance of disease and on the asymptomatic body respectively. While these processes are typical of the molecular era in biomedicine more generally, the array of new materials, structures, and devices, which nanomedicine makes envisageable, has the potential to intensify them significantly.

Consequently, an important implication of the predictive reasoning underlying nanodiagnostics is the manner in which it becomes entangled with conceptions of the body that contribute to blurring the classical line between the "normal" body and the "pathological" body. It creates a new state of existential uncertainty and a new medical figure: the "body-at-risk" of the "patient-in-waiting." Defined as "an umbrella

concept for those under medical surveillance between health and disease" [51] (p. 408), the patient-in-waiting is the direct consequence of the asymptomatic condition of potential disease generated by molecular diagnosis. Traditionally, medicine's field of activity has been marked by the pathological body [52]. In contrast, the focus of biomedicine is on a body-at-risk, a virtual always potentially pathological body produced by the predictive and probabilistic logic of nanodiagnostic devices. Nanodevices generate the molecular information that gives shape to this virtual state between health and disease. Consequently, it contributes to reversing the significance of the presence/absence of pathology that is, or had been, at the core of clinical medicine. Accordingly, within the emerging nanomedicine paradigm, the absence of pathology in the present is no longer an indicator of health. To take an example,

> If large-scale biobank research actually does identify "biomarkers" related to the onset or development of specific diseases, these biomarkers will become part of the definitions used for specific diseases, either broadening or differentiating them. An individual receiving a positive result on a micro-array for heart disease has this heart disease, whereas until recently his General Practitioner would have sent him away with the reassurance that nothing was wrong (or more precisely: that nothing could be found [32] (p. 16).

Nanodevices can be understood as "intermediary objects" [53], that is to say artifacts or technical vectors that materialize modes of molecular predictive reasoning, contributing to generating the body-at-risk as the new normal state of the body, and thus more broadly potentially transforming how we see and understand bodies. Rather than understanding our asymptomatic bodies as "healthy," we might begin to experience them as always being in a state of potentially becoming diseased. The normative power of molecular predictive knowledge, that is, the conception of health as a state of insecurity requiring perpetual surveillance, is supported by the emerging nanotechnology-based diagnostic technologies, increasingly able to process and analyze prodigious amounts of biological information without the direct intervention of a clinician or a physician.

The spectacular informational capacities of the aforementioned devices can produce a "hyperreal" understanding of health, to use the seminal concept of the sociologist Jean Baudrillard [54]. This is achieved by creating a simulated state in which the actual health status merges with the predictions generated by diagnostic devices. In other words, the simulated informational reality of the body, as it is projected probabilistically into the future, is experienced as the reality of the body in the present. The body-at-risk has the potential of shifting from being an exceptional condition to that of a quotidian reality. Moreover, the "rapidly advancing ability to bring potential futures into the present means that in theory each of us is now constituted as part of a single population, that of the pre-symptomatically ill, because we all are susceptible to one condition or another" [55] (p. 331). Indeed, with the increasing use of molecular diagnostic tools, the number of potential presymptomatic patients could grow considerably. The presymptomatic condition could foreseeably become the new norm for health. The capacity "to bring potential futures into

the present" could contribute to making the normal biological condition equivalent to that of the chronically ill, leading to a life of never ending medical surveillance [56].

Implantable wireless sensors [57] are an excellent example of not only nanomedicine's move toward the intensification of health surveillance [58] but also some further unforeseen implications associated with the pervasive nanocomputing paradigm applied in the context of health. Until fairly recently, health-monitoring procedures were for the most part restricted to the hospital or the clinic. Both of these institutional spaces accommodated relationships of intimacy and privacy between the patient and the clinician where clinical results could be given meaning. The nanotechnologically enabled miniaturization of devices could make possible a heightened form of bodily and health monitoring outside of both the clinic and the hospital. As it becomes feasible to wirelessly transmit information regarding slight physiological and molecular changes [59], the ensuing technological decentralization of medical monitoring puts into question the boundaries of corporal privacy and intimacy [60]. It equally poses questions regarding who will read the output, and how it will be made meaningful to both clinicians and patients [32]. Thus, new nanodevices of risk management are intertwined with the potential of contributing to transforming the intimate space of the individual body by opening it up as a broad space for technical intervention. In *The Vanishing Physician-Scientist?* Andrew Schaffer argues that increasingly "a widening schism between basic science and clinical medicine" is developing, imperiling communication and understanding between the two domains [61] (p. 7). In this context, due to their advanced molecular architectures and complex probabilistic modeling, nanodevices equally have the potential to generate an information intensive rendering of the body that might be difficultly legible to physicians let alone patients.

At the core of the emerging nanodiagnostic paradigm is the powerful notion of the desirability as well as the ability to be ahead of the disease. New devices capable of higher levels of precision and enhanced pervasive computing power will be able to detect potential dangers to health long before traditional symptoms become manifest, but perhaps also without the certainty that the predicted disease will eventually materialize. While acting earlier in disease prevention is certainly desirable, the exponential growth in the quantity of biomarkers and their associations with future diseases become, as we have seen, entangled with new conceptions of disease, health, and the body such as the "body-at-risk" and the "patient-in-waiting." These new modes of thinking about health and illness in turn invite the need for more pervasive forms of monitoring, further projecting our conception of health away from the present state of the body into a probabilistic future. In the conclusion we return to these implications; we now turn to another key dimension of the emerging nanomedicine paradigm, personalized medicine.

16.6 NANOMEDICINE AS PERSONALIZED MEDICINE

For many commentators, nanomedicine embodies a "patient-friendly" medical approach [62], bringing together both personalized diagnosis and therapeutic intervention, tailored to individual needs [7, 63]. Underpinning nanomedicine's privileged

position with respect to personalized medicine are developments in pharmacogenomics, pharmacogenitics, pharmacometabonomics, and pharmacoproteomics that have enabled scientists to make

> great strides in mapping the molecular pathways by which a change or mutation in a gene actually manifests itself as a disease. These advances have enabled drug researchers to develop diagnostic tools that can distinguish the subtypes of what had been considered a single disease, as well as chemical agents that target each [64] (p. 111).

Further developments in the field are potentially facilitated by nanomaterial devices optimized for biolabeling, molecular imaging, and improved drug discovery and delivery [65–68]. Prior to the emergence of molecular medicine, health was evaluated according to the average state of health of a population. The conception of health or disease was defined on a population-based logic. This had important consequences in terms of how diseases were identified and patients treated. Encapsulated in the vignette that follows as "trial-and-error medicine," critics argue that this is a general population model that must be overcome.

> A patient presents with symptoms, and the doctor makes a "mostly likely" diagnosis that is consistent with those symptoms, then prescribes a drug and, possibly, other treatment such as surgery. The drug dosage is typically based on the patient's weight. If the drug doesn't work or has significant side effects, the doctor may change the dosage or try another drug if one is available. Alternatively, the doctor may abandon the original diagnosis in favour of another and write a new prescription. The cycle is repeated until the correct, or more precise, diagnosis and treatment plan are discovered [64] (p.110).

In contrast, new molecular tools promise the ability to redefine "the boundary between normal and abnormal on a highly individual level" [32] (p. 13). However, it is important to understand that while several popular discourses suggest otherwise, molecular technologies lead to a particular type of "individualization": one that drills down into molecular causal pathways, and thus moves away from the patient as an individual. It does not, in fact, fundamentally depart from the population model. Rather, it is a refinement of the latter, based on a better stratification of patient groups according to their genetic information and biological variability. Nanomedicine's version of personalized medicine, as it is currently unfolding, would seem to contribute less to developing a type of medicine whose reference point is the individual him or herself than to a better categorization of molecular profiles, optimizing therapies suited to different population subgroups, and helping to reduce both under- and overtreatment.

Therefore, because the individual remains a case in a broader series of molecular categories, the emerging notion of "personalization" is better characterized as referring to a situation where it is "as if" the diagnosis were tailored for individual needs and the patient were a unique case [69]. Personalization operationalizes "individuality" by grouping individuals who share diagnostically or therapeutically relevant biomarkers. An important repercussion, but equally a formidable obstacle, is that the new "tailored" treatments challenge the blockbuster drug model of the pharmaceutical

industry [64]. However, if it is true that the personalized paradigm is poised to challenge the "one-size-fits-all" drug model by focusing on the molecular specificity of the patient, it is still a specificity produced by "refined" molecular categories.

Far from taking into account what Georges Canguilhem referred to as the individual and qualitative reality of the disease [52], the personalized medicine paradigm has the potential to heighten processes where the objectification of the patient's body leaves little room for the subjective experience of the disease. The patient her- or himself disappears as the focus is turned to biomarkers and molecular signatures. Thus, nanomedicine's personalized approach confronts us with a paradox: adapting medical practice to the unique case of each patient means focusing on what that patient shares with a molecularly specified subpopulation, thus abstracting away from his or her lived experience.

The import of how nanomedicine's paradigm of personalization might contribute to further decentering the subjective experience of the patient has to be understood in the context of a considerable body of scholarship in the social science and the humanities that have explored the significance of the lived experience of illness [70–74]. Researchers have documented the fact that illness is not merely experienced as a physiological pathology; it also "leads patients and those close to them to ask questions about what is happening, and to call into question ordinary explanations"; it provokes a "quest for meaning – Why me? Why now?" and can, in some instances, provoke biographical disruptions that unsettles a patient's sense of self [74] (p. 8). Although there can be no doubt that medicine's legitimacy is based on its technical efficacy, on its ability to detect and cure disease by objectifying pathological processes, it cannot be forgotten that "Illness combines physical and existential dimensions, bodily infirmity and human suffering" [72] (p. 70). As a result, "The hospital is not only the site of the construction and treatment of the medicalized body, but the site of a moral drama […] of human suffering and fear, of the confrontation with illness and death on the part of both the sick and those charged with their care" [72] (p. 85). This means that no matter the technical acumen of medicine, it always has to deal with the "irruption of the fundamentally moral dimensions of illness" into its "rational technical sphere" [72] (p. 85). Thus, in the contemporary setting, it is ineluctable that medicine has a broader moral dimension insofar as it provides "an understanding of the nature of suffering, and means of transforming and transcending suffering" [72] (p. 70). Medicine's ability to do so depends on the existence of skills and procedures that allow clinicians to tap into this broader experience of illness even as they perfect the tools of technical intervention to diagnose and treat disease.

Traces of the import of this dimension of medicine can be detected unevenly in the curricula of medical schools:

> every medical school strives to teach students a set of practices complementing the standard diagnostic and therapeutic activities. Courses in social medicine or the medical humanities and behavioural sciences teach students to attend to patient narratives and experience, to evaluate ethical issues in medical practice, and to consider the social context of illness and care. Forms of interviewing and assessing patients appropriate to

> such perspectives are also taught. In some curricula these issues receive scant attention; in some, such as the Patient-Doctor course of Harvard's New Pathway curriculum, a great deal of effort is devoted to teaching these complementary practices and the knowledge associated with this view of sickness and care [72] (p. 84).

It is worth noting that such efforts date to the beginning of the twentieth century and the introduction of natural-science-based education in medical schools. As early as 1910, the *Flexner Report on Medical Education in the United States and Canada* drew attention to the humanistic deficit that might be entailed by the institutionalization of science-based curricula in medicine [75]. Since then, efforts have been made to reduce the perceived shortfall through the introduction in curricula what are alternatively called "behavioral," "social," "psychological," "humanistic," and "ethical" components of health and illness, of "knowing patients as persons" [76]. An exploration of what represents the gold standard of personalization within the emerging nanomedicine paradigm, theranostics (therapeutics + diagnostics), draws attention to the prospect that "individualization" might move practioners further away from "knowing patients as persons."

Theranostic nanomedicine is a new strategy for the development of more specific personalized therapies to treat various diseases by combining diagnostic and therapeutic functionalities in a single biocompatible nanoagent or nanoplaform [63, 77–79]. Said nanoplatforms are expected to be able to diagnose disease, aim the real-time release of drugs at a targeted site, and monitor and evaluate treatment efficacy. These theranostic devices are the most advanced application of the concept of "intelligent nanomaterials" [80], which due to their physicochemical properties, are expected to "travel" easily into the body, overcoming its biological defense mechanisms [77]. They are designed to be versatile and effective "molecular missiles," with few side effects [81].

As Bensaude-Vincent and Loewe have noted, warfare metaphors such as "therapeutic missiles," "nano-weapons," or "smart bombs" capable of precision targeting thus avoiding "collateral damage" are commonly used to convey the capacity for control, surveillance, and efficiency of these nanodevices [82]. These devices give a whole new meaning to Paul Ehrich's popular image of the "magic bullet," which is extensively used in articles mining the potentialities of nanomedicine. "Magic bullets" are in fact rare "in therapeutic reality because it is challenging to find particular targets among the trillions of cells in a typical body and because the bullet must overcome a number of barriers to get inside the target cell" [83] (p. 544). Beyond the debate about the "therapeutic reality" that also includes notable concerns regarding "toxilogical issues" [63] (p. 1412); [46] (p. 23), warfare metaphors convey a decentering of the individual experience of illness. This is because the individual's body is framed as the terrain on which the war is fought. Nanodevices "reconnoiter" the molecular landscape returning "actionable intelligence" that leads to the "launch" of a therapeutic action. The mediation of the therapeutic relationship via devices autonomously communicating information and provoking responses implies a doctor–patient relationship that runs the risk of being completely technologically driven.

This potential implication comes into clear focus if we explore the ideal practioners of nanomedicine as envisioned by those who most energetically promote a nano-enabled theranostics. Earlier, we saw how Aspinall and Hamermesh identify the limits inherent in "trial-and-error-medicine." Transitioning toward a model of personalized medicine requires not only important transformations in the pharmaceutical industry and regulatory frameworks but also changes in the training and habits of the medical profession [64]. More specifically, Vizirianakis and Amanatiadou argue that if physicians are to adequately take advantage of nanomedically enabled theranostics, they must become familiar with and be able to apply pharmagenomics knowledge. This requires "appropriate education […] through the development of new curricula and educational approaches. The training has to be focused on pharmacogenomics, personalized medicine, and pharmacotyping concepts as well as bioinformatics and information-based medical practice" [68] (p. 123). The expected outcome of such measures is a changed "drug-delivery environment" whose description is worth quoting in full:

> This means that the drug-delivery environment is changing from a drug-selection process where physicians use their own clinical experience into a more highly integrated, information-based and computer-aided process. Indeed, the development of the computer-based decision support systems in drug delivery represents an effort for the automation of prescription to achieve quick processing and allow the inter-correlation of multiple patient and drug-related factors for reducing errors compared to traditional handwritten prescriptions. However, in order for such a direction to gain full advantage worldwide, it is important for the computer-based decision systems to be designed, managed and updated in a way to ensure interoperability with other computerized methods for managing healthcare data and common standards between different national systems [68] (pp. 128–129).

This vision, whether realizable or not, clearly depicts a doctor–patient interaction where the gaze of the doctor is no longer on the patient as such. Instead, the "individuality" of the patient is materialized on the multiple screens that aggregate, correlate, and analyze molecular data. Indeed, in an ideal theranostic situation, the patient could be monitored at a distance; his or her physical presence would not be required. However even in the context of face-to-face interaction, the automation of therapeutic decisions, based on the pervasive computing of molecular data, potentially transforms the role of a physician to that of a technician.

The struggle to balance care with the objectivity that grounds the technological efficacy of curing is not a new one. As the French surgeon and historian Alain-Charles Masquelet notes, "the evolution of modern surgery can be read as the history of a double ebb: the withdrawal of the operator's hand [the surgeon] along with the closure of the suffering bodies" [84] (p. 139).[1] Since its inception, the gradual distancing of medical practitioners from patients via the use of technical instruments has underwritten the scientificity of the medical act. This distancing could achieve a qualitatively new dimension as a result of the projected mediation of diagnostic and therapeutic functions via theranostic devices.

[1] Our translation: "L'évolution de la chirurgie moderne peut être lue comme l'histoire d'un double reflux: retrait de la main de l'opérateur en même temps que s'accomplit une clôture des corps souffrants."

Today, the representation of biomedical progress is linked to an ever-more technologically driven medical practice where the physician is less and less directly involved. The focus on new noninvasive devices like theranostic nanoagents, which require neither an intrusion into the body through skin incision nor direct human action, rests on the conviction of the scientific superiority of "technomedicine" [85–87]. It would be unwise to reject the promise associated with these new technologies; they certainly contain the potential to treat, manage, and even cure many debilitating diseases. Nonetheless, however successful theranostics might come to be, not all disease will be curable and death will continue to mark the ultimate horizon of our embodied existence. Even in the case of success, the biomedical experience will continue to have an existential dimension for the patient that analysis of molecular pathways cannot convey. The moral dimension of a physician's role arises from the fact he or she is tasked with mediating between the lived experience of his or her patient's illness and the formidable technical knowledge of modern medical practice. As such physicians occupy and manage a trading-zone where the frailty and fear associated with illness is converted into the hope associated with curing and perhaps postponing death. Physicians' ability to do so rests on their ability to clearly grasp the technological potential and limits associated with the advance of medicine while never losing sight of the patient as a person.

16.7 NANOMEDICINE AS REGENERATIVE MEDICINE

Regenerative medicine is certainly the fullest expression of the engineering logic underlying the nanomedical paradigm captured by the "machine of life" and "programmable matter" or "life as technology" metaphors [87]. Regenerative medicine emerges from "applied stem cell and developmental biology" [88] (p. 773). Its goal is to artificially regenerate human cells, tissue or organs to restore normal functioning using "somatic, adult stem or embryo-derived cells," and more recently even embryo cells "reprogrammed from adult cells" [89] (p. 1). What nanotechnology promises to bring to the clinical treatment of tissue degeneration is a bioengineering therapeutic approach. In the two previous sections we have drawn attention to the manner in which nanomedicine's revolutionary irruption in biomedicine can in many instances be understood in terms of the intensification of existing trends both in the molecular medicine model, or indeed going back further, as a result of the tension between technical efficacy and humanistic engagement that has been at the heart of modern medicine. It is in the context of regenerative medicine, however, where nanomedicine has the potential to open up radically new horizons for biomedicine and perhaps dramatically change how we think about health, illness, and bodies. However, as a result of the infancy of the field, but also of its potential novelty, it is appreciably more challenging to understand or begin to work out the implications of this area of the emerging nanomedicine paradigm. Consequently, in this last section we comment briefly on a central paradox of regenerative medicine as it is currently developing.

"The focus of regenerative medicine is to work with the body's own repair mechanisms" in order to "help the body to heal itself" [7] (p. 9 and 27). It aims "to strengthen the self-healing processes of the human body either by stimulating or emulating them" [90] (p. 193). The salamander's regenerative potential is frequently the reference model [91, 92]. Consequently, regenerative medicine differs significantly from its predecessor transplantation medicine in that it doesn't intend to "replace" or "repair," but to "regenerate" biological materials by controlling the regenerative process. Whereas transplantation medicine aims to overcome the degenerative process by replacing an organ, regenerative medicine seeks to reestablish the organ's "normal" structure and function without substitution. By stimulating or emulating auto-regeneration, it attempts not only to stop but also to "reverse" the tissue degeneration in order to recover a former, or "original" condition. In other words, regenerative medicine aims to mobilize bioengineering methods that seek to heal by reestablishing the physiological normality of a tissue that exists before disease occurs. In this sense, it aims not to be "ahead of" but "before" the disease.

Nonetheless, the notion of returning to a former "normal" physiological condition is paradoxical in relation to the actual engineering approach pursued through regenerative medicine. In reality, it does not heal by "reestablishing" a past biological condition but by "re-shaping" the present through the use of biomaterials or cellular cultures. The expression to "help the body to heal itself" is highly representative of this paradox. Regenerative medicine, as it is envisioned within the emerging nanomedicine paradigm, in contrast with organ transplantation, aims not to restore but instead to reshape the body, giving a new direction to biological development itself [93]. If organ transplantation is based on a mechanical and fragmented representation of the human body, regenerative medicine is based on a re-shapable, re-programmable representation of the body where the border between the natural and the artificial becomes blurred [56]. This conception borrows from molecular biology and biotechnology the idea of the fundamental "plasticity" or "artificiality" of life, and adds the building logic of the engineer to produce a vision of life and the body as synthesizable, reproducible, malleable and transformable material. How will this envisioned openness of the human body to nano-engineered change become entangled with new conceptions of the body, health, and illness? This is a question that cannot be answered in the abstract. We must wait until or when this potential is materialized in concrete technologies. As a result, the question remains for the moment a rhetorical one. However, it is perhaps one that is worth repeating as the technology matures because some day it might cease being so, because we will have the ability and need to attempt to produce an answer.

16.8 CONCLUSION

In this chapter, we have argued that the absence of the more radical transformations predicted by nanomedicine should be read as neither indicating the failure of the nanomedicine paradigm nor requiring us to wait until the predicted transformations actually occur before trying to work out their implications. Instead, we have argued

that the initial extrapolations, anchored in a technological determinist logic as they were, have to be understood less as a roadmap of what was or is to come, and more as a rallying cry to focus resources and effort. In this latter sense there can be no doubt the nanomedicine has succeeded. It is currently a successful research program and though actual clinical applications have to date been few, there is, however, at present no plausible reason to believe that they will not come in the future. Developing an understanding of the implications of such changes requires researchers, policy makers, and stakeholders to frame the analysis of technological development in its contextual nature: the manner in which technologies develop is highly contingent, depending on both local conditions in the laboratory and broader social, cultural, political, and economic dynamics. As a contribution to this type of exploration, we have proposed an analysis of how nanomedicine is or can become entangled and intertwined with conceptions of health, illness, and the body.

We have also argued that searching for moments of radical breaks perhaps prevents us from actually understanding the real time implications associated with nanomedicine as it is currently developing. It prevents us from grasping how continuities can also be productive of change: "emerging biomedical technologies are often likely to shift or reinterpret the goals of medicine, even if they pretend to be just more effective means towards well known and widely accepted goals" [32] (p. 16). This is the reason why we have argued that a fruitful avenue for trying to grasp the implications of nanomedicine is by understanding its continuity with the molecular medicine paradigm. This paradigm is particularly significant. As a result of its bioengineering logic, it has enabled nanomedicine to project itself into three key areas of contemporary biomedicine: predictive, personalized, and regenerative medicine.

In the case of the first area, predictive medicine, we have shown how developments in the field have the potential of becoming enmeshed with new forms of understanding health, disease, and the body conveyed by expressions such as the "body-at-risk" and the "patient-in-waiting," deriving from the paradox that successful disease detection is associated with the absence rather than the presence of actual pathology. New nanomaterials, devices, and computing architectures greatly intensify the ability to more complexly model molecular-based disease pathways. While the temporal logic of being "ahead" of the disease is reasonable, should this understanding of health become entangled more broadly within medical practice, we run the risk of a temporal treadmill where researchers and health systems searching for ever earlier signs of possible disease formation will create the need for new forms of intensive and extensive surveillance. With respect to personalized medicine, we have shown that again despite appearances to the contrary the nanomedicine paradigm does not represent so much a break with population-based health, but its molecular refinement. The model of predictive medicine developing within the nanomedicine can also be understood in terms of a second continuity, one that is far reaching and coextensive with modern medicine itself, namely, the struggle to balance objectivity, through the use of diagnostic and therapeutic technologies, with an understanding of the patient's existential experience of disease. Paradoxically, nanomedicine-based predictive medicine seems best able to capture the individuality of the patient when it abstracts from his or her lived experience. And in the case of theranostics, it even

seems that the judgment of the clinician him or herself runs the risk of being displaced by the command and control functions of pervasive computing at the molecular level. However successful theranostics can or shall be, the horizon of disease and death will remain with all their concomitant existential dilemmas and fears. Nanomedicine will need to find a way to make room for the patient as person. A nanomedically based regenerative medicine tests our ability to explore its implications, because of its infancy, but also because of its view of bodily development as being open-ended, conceiving the body as "programmable matter." The paradox in this case is that although regenerative medicine promises to "re-store" a previous state, in reality it is "re-shaping" the present and opening up new developmental possibilities in the future. What this entails cannot at present be answered, but it is a question that we must continue asking in "real time" as the technology develops. Indeed, in our attempt to identify implications, our purpose has not been to provide answers about nanomedicine's future, but rather to provide questions that we hope can contribute to developing the reflexivity required to understand nanomedicine's development.

REFERENCES

[1] Cientifica. Global Funding of Nanotechnologies and Its Impact. Available at: http://www.cientifica.com/research/white-papers/global-nanotechnology-funding-2011/. Accessed December 22, 2014.

[2] Berger M. Nano-Society: Pushing the Boundaries of Technology. London: Royal Society of Chemistry; 2009.

[3] Duncan R, Gaspar R. Nanomedicine(s) under the microscope. Molecular Pharmaceutics 2011; 8(6): 2101–2141.

[4] Venkatraman S. Has nanomedicine lived up to its promise? Nanotechnology 2014; 25(37): 372501.

[5] Etheridge ML, Campbell SA, Erdman AG, Haynes CL, Wolf SM, McCullough J. The big picture on nanomedicine: The state of investigational and approved nanomedicine products. Nanomedicine: Nanotechnology, Biology and Medicine 2013; 9(1): 1–14.

[6] Juliano R. The future of nanomedicine: Promises and limitations. Science and Public Policy 2012; 39(1): 99–104.

[7] European Technology Platform (ETP). Vision Paper and Basis for a Strategic Research Agenda for Nanomedicine. Luxembourg: Office for Official Publications of the European Communities; 2005.

[8] López J. Bridging the gaps: Science fiction in nanotechnology. HYLE—International Journal for Philosophy of Chemistry 2004; 10(2): 129–152.

[9] Hamilton SN. Traces of the future: Biotechnology, science fiction, and the media. Science Fiction Studies 2003; 30(2): 267–282.

[10] European Technological Platform in Nanomedicine. Nanomedicine 2020: Contribution of Nanomedicine to Horizon 2020. Paris: European Technology Platform for Nanomedicine (ETPN); 2013.

[11] Anderson A, Petersen A, Wilkinson C, Allan S. Nanotechnology, Risk and Communication. London: Palgrave Macmillan; 2009.

[12] Kulkarni RP. Nano-Bio-Genesis: Tracing the rise of nanotechnology and nanobiotechnology as 'big science'. Journal of Biomedical Discovery and Collaboration 2007; 2: 3.

[13] Gelfert A. Nanotechnology as ideology: Towards a critical theory of 'converging technologies'. Science Technology & Society 2012; 17(1): 143–164.

[14] McCray WP. Will small be beautiful? Making policies for our nanotech future. History and Technology 2005; 21(2): 177–203.

[15] Wood S, Jones R, Geldart A. The Social and Economic Challenges of Nanotechnology. Swindon: Economic and Social Research Council; 2003.

[16] Porter AL, Youtie J. Where does nanotechnology belong in the map of science? Nature Nanotechnology 2009; 4(9): 534–536.

[17] Sismondo S. An Introduction to Science and Technology Studies. New York: John Wiley & Sons; 2011.

[18] Williams R, Edge D. The social shaping of technology. Research Policy 1996; 25(6): 865–899.

[19] Guston DH, Sarewitz D. Real-time technology assessment. Technology in Society 2002; 24(1): 93–109.

[20] Rabinow P, Bennett G. Synthetic biology: Ethical ramifications 2009. Systems and Synthetic Biology 2009; 3(1–4): 99–108.

[21] Hanson VL. Envisioning ethical nanotechnology: The rhetorical role of visions in postponing societal and ethical implications research. Science as Culture 2011; 20(1): 1–36.

[22] López JJ. Nanotechnology: Legitimacy, narrative and emergent technologies. Sociology Compass 2008; 2(4): 1266–1286.

[23] Geels FW, Smit WA. Failed technology futures: Pitfalls and lessons from a historical survey. Futures 2000; 32(9): 867–885.

[24] Borup M, Brown N, Konrad K, Van Lente H. The sociology of expectations in science and technology. Technology Analysis & Strategic Management 2006; 18(3–4): 285–298.

[25] Crabu S. Nanomedicine in the making. Expectations, scientific narrations and materiality. TECNOSCIENZA: Italian Journal of Science & Technology Studies 2014; 5(1): 43–66.

[26] Guston DH. Understanding 'anticipatory governance'. Social Studies of Science 2014; 44(2): 218–242.

[27] Nordmann A. If and then: A critique of speculative nanoethics. Nanoethics 2007; 1(1): 31–46.

[28] Selin C, Hudson R. Envisioning nanotechnology: New media and future-oriented stakeholder dialogue. Technology in Society 2010; 32(3): 173–182.

[29] Noury M, Lafontaine C. De la nanomedicine à la nanosanté: vers un nouveau paradigme biomédical. Socio-Anthropologie 2014; 29: 13–36.

[30] Hoad TF. Concise Oxford English Dictionary of Etymology. Oxford: Oxford University Press; 1996.

[31] Noury M. La nanosanté. Perspectives et enjeux sociologiques de l'application des nanotechnologies à la médicine. Montreal/Nanterre: Université de Montréal/Université Paris Ouest; 2014.

[32] Boenink M. Molecular medicine and concepts of disease: The ethical value of a conceptual analysis of emerging biomedical technologies. Medicine, Health Care and Philosophy 2010; 13(1): 11–23.

[33] Rose N. The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton, NJ: Princeton University Press; 2007.

[34] Keller EF. Refiguring Life: Metaphors of Twentieth-Century Biology. New York: Columbia University Press; 1995.

[35] Keller EF, Keller EF. The Century of the Gene. Cambridge, MA: Harvard University Press; 2009.

[36] Kay LE. Who Wrote the Book of Life?: A History of the Genetic Code. Stanford, CA: Stanford University Press; 2000.

[37] Boisseau P, Loubaton B. Nanomedicine, nanotechnology in medicine. Comptes Rendus Physique 2011; 12(7): 620–636.

[38] Jain KK. Nanomedicine: Application of nanobiotechnology in medical practice. Medical Principles and Practice 2008; 17(2): 89.

[39] Burgess R. Understanding Nanomedicine: An Introductory Textbook. Hoboken, NJ: Pan Stanford; 2012.

[40] Mohandy C, Arya G, Verma RS, Sahoo SK Nanobiotechnology: Application of nanotechnology in therapeutics and diagnosis. International Journal of Green Nanotechnology: Biomedicine 2009; 1: B24–B38.

[41] Rothman BK. Genetic Maps and Human Imagination. New York/London: W.W. Norton; 1998.

[42] National Science and Technology Council (NSTC). Nanotechnology: Shaping the World Atom by Atom. Washington, DC: National Science and Technology Council; 1999.

[43] Thacker E. Biomedia. Minneapolis, MN: University of Minnesota Press; 2004.

[44] Jain KK. Biomarkers: Technologies, Markets and Companies. Dublin: Research and Markets; 2007.

[45] Lock M. The future is now: Locating biomarkers for dementia. In: Burri RV, Dumit J, eds. Biomedicine as Culture: Instrumental Practices, Technoscientific Knowledge, and New Modes of Live. New York: Routledge; 2007, pp. 61–86.

[46] Logoththetidis S. Nanomediciine: The medicine of tomorrow. In: Logotheidis S, ed. Nanomedicine and Nanotechnology. Berlin/Heidelberg: Verlag/Springer; 2012, pp. 1–26.

[47] Syed MA. Advances in nanodiagnostic techniques for microbial agents. Biosensors and Bioelectronics 2014; 51: 391–400.

[48] Dürr A, Herson A, Rosenblum O, Gargiulo M. Enjeux de la médecine prédictive en neurogénétique: quelle place pour l'homme? Champ psy 2010; 55(3): 83–95.

[49] Reynaudi M, Sauneron S. Médicine prédictive: les balbutiements d'in concept aux enjur considerables. Available at: http://www.cientifica.com/research/white-papers/global-nanotechnology-funding-2011/. Accessed December 22, 2014.

[50] Hood L. A Vision for Personalized Medicine. Available at: http://www.technologyreview.com/news/417929/a-vision-for-personalized-medicine/. Accessed December 22, 2014.

[51] Timmermans S, Buchbinder M. Patients-in-waiting: Living between sickness and health in the genomics era. Journal of Health and Social Behavior 2010; 51(4): 408–423.

[52] Canguilhem G. The Normal and the Pathological. New York: Zone Books; 1991.

[53] Vinck D. Accessing Material Culture by Following Intermediary Objects. Rijeka: European Union; 2012.

[54] Baudrillard J. Simulacra and Simulation. Ann Arbor, MI: University of Michigan Press; 1995.

[55] Lock M, Nguyen V. An Anthropology of Biomedicine. Chichester: Wiley-Blackwell; 2010.

[56] Lafontaine C. La société post-mortelle. Paris: Seuil 2008.

[57] Bashirullah R. Wireless implants. Microwave Magazine, IEEE 2010; 11[7): S14–S23.

[58] Gordon N, Sagman U. Nanomedicine Taxonomy. 2003; Available at: http://www.ifm.liu.se/physchem/courses/NKEC18/Nanomedicinetaxonomy.pdf. Accessed December 22, 2014.

[59] Walhout B, van Keulen I, van Est R, Malsch I. Nanomedicine in the Netherlands: Social and Economic Challenges. The Hague: Rathenau Institute; 2010.

[60] Lafontaine C. Nouvelles technologies et subjectivité: Les frontières renversées de l'intimité. Sociologie et Societes 2003; 35(2): 203–212.

[61] Schaffer A. The Vanishing Physician Scientists? New York: Cornell University Press; 2009.

[62] Ranganathan R, Madanmohan S, Kesavan A, Baskar G, Krishnamoorthy YR, Santosham R, et al. Nanomedicine: Towards development of patient-friendly drug-delivery systems for oncological applications. International Journal of Nanomedicine 2012; 7: 1043–1060.

[63] Mura S, Couvreur P. Nanotheranostics for personalized medicine. Advanced Drug Delivery Reviews 2012; 64(13): 1394–1416.

[64] Aspinall MG, Hamermesh RG. Realizing the promise of personalized medicine. Harvard Business Review 2007; 85(10): 108–117.

[65] Jain KK. Personalized medicine. Current Opinion in Molecular Therapeutics 2002; 4(6): 548–558.

[66] Jain K. Role of pharmacoproteomics in the development of personalized medicine. Pharmacogenomics 2004; 5(3): 331–336.

[67] Jain KK. Role of nanobiotechnology in developing personalized medicine for cancer. Technology in Cancer Research & Treatment 2005; 4(6): 645–650.

[68] Vizirianakis IS, Amanatiadou EP. Pharmacogenomics and Nanotechnology: Towards advancing presonalized medicine. In: Logoththetidis S, ed. Nanomedicine and Nanobiotechnology. Berlin/Heidelberg: Verlag/Springer; 2012, pp. 115–134.

[69] Ricroch A, Dekeuwer C. Enjeux éthiques de la médecine prédictive. Droit, déontologie & soin 2007; 7(2): 169–194.

[70] Bury M. The sociology of chronic illness: a review of research and prospects. Sociology of Health & Illness 1991; 13(4): 451–468.

[71] Charmaz K. Experiencing chronic illness. In: Albrecht GL, Fitzpatrick R, Scrimshaw SC, eds. The Handbook of Social Studies in Health and Medicine. London: Sage; 2000, pp. 277–292.

[72] Good B. Medicine, Rationality and Experience: An Anthropological Perspective. Cambridge, UK: Cambridge University Press; 1994.

[73] Williams S. Chronic illness as biographical disruption or biographical disruption as chronic illness? Reflections on a core concept. Sociology of Health & Illness 2000; 22(1): 40–67.

[74] Pierret J. The illness experience: State of knowledge and perspectives for research. Sociology of Health & Illness 2003; 25(3): 4–22.

[75] López J. How sociology can save bioethics… maybe. Sociology of Health & Illness 2004; 26(7): 875–896.

[76] Fox RC. Is medical education asking too much of bioethics? Daedalus 1999; 128: 1–25.

[77] Jeelani S, Reddy RJ, Maheswaran T, Asokan G, Dany A, Anand B. Theranostics: A treasured tailor for tomorrow. Journal of Pharmacy & Bioallied Sciences 2014; 6(Suppl 1): S6.

[78] Pan D. Theranostic nanomedicine with functional nanoarchitecture. Molecular Pharmaceutics 2013; 10(3): 781–782.

[79] Prabhu P, Patravale V. The upcoming field of theranostic nanomedicine: An overview. Journal of Biomedical Nanotechnology 2012; 8(6): 859–882.

[80] Lehner R, Wang X, Marsch S, Hunziker P. Intelligent nanomaterials for medicine: Carrier platforms and targeting strategies in the context of clinical application. Nanomedicine: Nanotechnology, Biology and Medicine 2013; 9(6): 742–757.

[81] Patel S, Bhirde AA, Rusling JF, Chen X, Gutkind JS, Patel V. Nano delivers big: Designing molecular missiles for cancer therapeutics. Pharmaceutics 2011; 3(1): 34–52.

[82] Bensaude VB, Loeve S. Metaphors in nanomedicine: The case of targeted drug delivery. Nanoethics 2014; 8(1): 1–17.

[83] Plank C. Nanomedicine: Silence the target. Nature Nanotechnology 2009; 4(9): 544–545.

[84] Masquelet AC. La relégation du corps à corps chirrgical. In: le corps relègué. Masquelet AC, ed. Paris: PUF; 2007, pp. 139–149.

[85] Pickstone JV. Ways of knowing: towards a historical sociology of science, technology and medicine. The British Journal for the History of Science 1993; 26(4): 433–458.

[86] Pickstone JV. Ways of Knowing: A New History of Science, Technology, and Medicine. Chicago, IL: University of Chicago Press; 2001.

[87] Landecker H. Culturing Life. Cambridge, MA: Harvard University Press; 2007.

[88] Mironov V, Visconti R, Markwald R. What is regenerative medicine? Emergence of applied stem cell and developmental biology. Expert Opinion on Biological Therapy 2004; 4(6): 773–781.

[89] Mason C, Dunnill P. A brief definition of regenerative medicine. Regenerative Medicine 2008; 3(1): 1–5.

[90] Ach JS, Lüttenberg B. Nanobiotechnology, Nanomedicine and Human Enhancement. Munster: LIT Verlag; 2008.

[91] Brockes JP, Kumar A. Appendage regeneration in adult vertebrates and implications for regenerative medicine. Science 2005; 310(5756): 1919–1923.

[92] Kahn A, Papillon F. Le secret de la salamandre. La Médecine En Quête D'immortalité. Paris: éditionsNil; 2005.

[93] Hogle LF. Life/time warranty: Rechargeable cells and extendable lives. In: Lock M, Franklin S, eds. Remaking Life and Death: Toward an Anthropology of Biosciences. Santa Fe, NM: School of American Research Press; 2003, pp. 61–96.

PART IV

ETHICAL AND LEGAL ASPECTS OF WIRELESS COMPUTING IN MEDICINE

17

ETHICAL CHALLENGES OF UBIQUITOUS HEALTH CARE*

WILLIAM SIMS BAINBRIDGE

Division of Information and Intelligent Systems, National Science Foundation, Arlington, VA, USA

The views expressed in this essay do not necessarily represent the views of the National Science Foundation or the United States

17.1 INTRODUCTION

While there are many dimensions of medical ethics, the one that has received the most discussion in the context of the new information and communications technologies is privacy. However, much of the debate has not been specific to medicine, and often focused on narrow security issues [1]. Computer scientists often define privacy as security, seeking infallible technical methods to prevent intruders from gaining access to private data, and yet decades of work have failed to provide perfect security. For example, anonymization is a method to allow analysis of data such as those associated with public health, without revealing the identity of individuals, yet it often fails to disguise people's identities effectively [2]. At the risk of oversimplifying one of humanity's most complex sets of issues, it is worth noting that much of the social scientific discussion of privacy occurred as early as the 1960s, which was a time of cultural turmoil as well as rapid development of such fields as sociology of medicine.

* For *Pervasive Computing in Medicine*.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

Perhaps the most illuminating relevant publication of that period was "The Social Psychology of Privacy," by Barry Schwartz, published in 1968 in the *American Journal of Sociology* [3]. Today, publications about privacy tend to start from a situation in which personal data that an individual wishes to keep private are vulnerable to disclosure for one or another technical reason. This narrow conception fails to provide any deep understanding of the human motivations to protect or violate privacy, and despite seeming "modern" is quite insufficient to guide us in the age of the Internet. Schwartz, so long ago, offered the needed fundamental insight: Society is a dynamic system, characterized by both social cohesion (intimacy) and social separation (privacy). He noted that social status divisions in society require and impose privacy, as the societal elite withdraws from the common people, while possessing considerable power to violate the privacy of ordinary citizens. Kinship groups also create divisions in the social structure, across which many kinds of information do not flow easily, thus providing a degree of privacy for families.

1968 America—the environment in which Schwartz published—was a time of great social conflict symbolized by the assassinations of Martin Luther King, Jr., and Bobby Kennedy, the Tet Offensive in the Vietnam War, and protests at the Democratic National Convention in Chicago. Even social scientists whose own values were quite conventional subscribed to a *conflict theory* of human society, in which competition between social classes, religions, ethnic groups, political ideologies, and criminal associations were inevitably endemic. Under such circumstances, a universally accepted system of ethics, supporting privacy among other goals, would be impossible. In subsequent decades, this 1960s' view of society as a battleground faded from popularity, yet seems to have returned in the second decade of the subsequent century. While conflict theory cannot be dismissed, we cannot state with any scientific certainty how correct or incorrect this theory is, merely begin with the observation that it renders general moral systems problematic.

One theory of ethics conceptualizes it merely as conformity to the norms of the given society [4]. A subset of that theory imagines a pure democracy in which members of the society have equal power to determine what those norms should be, presumably not merely serving their own economic interests but expressing through their votes some sympathy for the needs of other people. Yet the standard social scientific model of power in society is the Iron Law of Oligarchy, that some kind of elite class always has greater power, even as the society may oscillate between two unsustainable extreme conditions: democracy and tyranny [5, 6].

Many traditional societies imagined that fundamental norms and values were god-given, perhaps in the form of the Ten Commandments received on a mountaintop, and even today many people look to ancient traditions for their moral guidance. In modern secular societies, informal norms become established through legislation, lacking any supernatural justification. Morality is nothing more than public policy. Periodically, societal leaders offer programs to improve public morality, and recently one of their areas of attention has been privacy. In February 2012, the federal US administration issued a Consumer Privacy Bill of Rights, asserting seven principles that private companies should respect, but that logically would apply as well to government agencies, and to healthcare providers; however, they were legally established.

1. Individual control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
2. Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
3. Respect for context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
4. Security: Consumers have a right to secure and responsible handling of personal data.
5. Access and accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data are inaccurate.
6. Focused collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
7. Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights [7].

Each one of these seven principles uses the word *right*, yet the fundamental source of rights remains controversial. Did the American people inherit the "divine right of kings" when they deposed their English monarch in 1776? Or is the word *right* merely rhetoric to promote a law that has not been formally legislated? At the same time, a vigorous debate is raging, in politics and the popular press, about government intrusion in citizens' privacy under the justification of national security. A topic related to privacy has a much better intellectual foundation: intellectual property.

Patents, copyrights, trademarks, trade secrets, and the like have been the subject of debate and analysis for many years, although the Internet has changed the technological basis on which their stability rests [8]. Privacy may be conceptualized as one more category of intellectual property: information to which a person or group wishes to retain exclusive use. Violations of individual privacy stir up public emotions, but theft of valuable information that is not personally embarrassing can cause equal harm. In ancient days, notions of property varied across societies, most obviously in the ownership of land, and inheritance practices [9, 10]. Was land owned by individuals, families, or the community as a whole? If an individual is the owner, do all the children inherit upon the owner's death, or only the first-born son? Applied to privacy, with the insights of Schwartz in mind, do members of the primary social group have no privacy among themselves, but great privacy with respect to outsiders? Considering the ownership of intellectual products, ancient writers like Sophocles and Homer were not considered plagiarists for building upon the works of earlier authors. Are copyrights a temporary custom, technologically obsolete now that anyone may copy text, music, and video at minimal cost?

Recognizing the connection between privacy and intellectual property allows us to see a broader range of ethical issues in the age of Internet. For doctors and patients to make the optimal decisions about diagnosis, treatment, and prognosis, would it not be necessary to place all scientific medical publications in an online digital library,

freely available to everyone? If this seems like a good idea, how will the journals be supported in their difficult work of reviewing and publishing scientific papers if they cannot charge subscription fees? Should every pharmaceutical company be required to immediately post online all findings from its internal research and development efforts, the better to speed the progress of medical advance? Should every medical patient contribute full case history and test data to a central archive to assess disorder causes and treatment effectiveness through systematic epidemiology? The last question implies that private, personal data would be widely shared.

While provocative, these questions are also naive, and each triggers endless arguments. Indeed, debates over the ownership of information have raged for decades, and have only expanded in recent years with the advance of information technology [11, 12]. At present, the National Institutes of Health in the United States require that all journal articles based on research they funded be accessible online through PubMed Central [13]. But this policy does not apply to journal articles based on the research funded by other sources. Suppose we scan into our computer medical publications that are covered by copyright and are not in PubMed, then post them on our website as a contribution to human well-being. Have we behaved ethically or unethically?

17.2 A PHILOSOPHICAL FRAMEWORK

The example of privacy allows us to sketch some general issues that then can be explored in the context of other ethical realms. The classical way to conceptualize medical ethics is to frame these as a coherent set of principles formally established to guide the actions of the members of a particular well-defined profession. From the perspective of the philosophy of ethics, this is a very practical approach, connected to the traditional rationalist view that ethics can be based in a *social contract* rather than requiring some transcendental source like the commandments of God [14, 15]. However, many modern schools of philosophy raise doubts about whether ethics can ever have a solid basis that renders morality more than merely the customs of a particular group [16, 17].

Often, a single organization seeks to speak for a profession, as when the American Medical Association promulgated a nine-point Code of Ethics, beginning thus: "A physician shall be dedicated to providing competent medical care, with compassion and respect for human dignity and rights" [18]. But the ethical issues associated with health are not limited to doctors, involving also medical schools, pharmaceutical companies, insurance systems, mass media, and indeed concerning the behavior of the patients themselves. We may start with four simple points as follows:

1. The rights of the patient are of obvious importance but somewhat difficult to define precisely.
2. Medical professionals also have rights, in addition to their duties.
3. Patients have duties as well as rights.
4. Many ethical implications of medical treatment extend outside the doctor's office, giving the wider society a complex pattern of rights and duties.

The second point reminds us that medical personnel are fallible human beings, and that patient records include considerable data about their behavior as well. Consider this real-life example. A patient is scheduled for minor surgery. Medical professional A administers a local anesthetic at noon. Medical professional B performs the surgery at 2 o'clock. The patient feels significant pain. Was B at fault either for doing the surgery later than scheduled, after the anesthetic had begun to wear off, or for failing to be alert to the patient's condition. Or was A at fault for failing to communicate effectively either with B or with the patient? Or should the patient expect that some amount of pain must be endured? Does the medical record give the times of the actions by the two medical professionals, and the patient's expression of pain? This example was selected because the answers to the questions are unclear, and the net result did not endanger the future health of the patient. Much more serious cases may also possess ambiguity.

For years, the norms concerning medical malpractice have been hotly debated, and great uncertainty exists on which public policy would really reduce it [19, 20]. Increased penalties may merely raise the cost of malpractice insurance, a cost ultimately passed on to the patient or the taxpayer [21]. No one suggests that doctors should be held blameless for their most egregious acts of incompetence, yet they are not infallible, and all kinds of accidents may occasionally happen, even under carefully controlled conditions. Part of the philosophical debate has concerned to what extent and in what way the doctor takes responsibility for the patient, rather than sharing responsibility with the patient [22]. In terms of information technology, the issue may become how aggressively the database records the behavior of the medical professional, and what information would be relevant for adjudication of malpractice claims.

Doctors, nurses, and patients are not the only actors in this drama, because they use tools, medications, and insurance systems created by other people, which may be faulty. The most famous example involving information technology is the Therac-25 case from the 1980s [23, 24]. The computer program that controlled this radiation therapy device contained a "bug," probably one that let different procedures get out of proper synchronization, disastrously overdosing six patients. This case alerted the field of computer-controlled medical devices to the issue of software reliability, and much progress has occurred since then. Yet perfect reliability has not been achieved, and much research has shown it is extremely difficult for computer programmers and specialists in other fields to cooperate effectively in producing new information systems [25, 26]. At the very least, they must be aware of their professional responsibilities, despite not being in the treatment room with the doctor and patient.

We do not ordinarily think of patients as having duties, although medical professionals are often frustrated by their patients' failure to follow directions, such as noncompliance taking prescribed medications. Yet a frequently cited analysis by Talcott Parsons, among the most influential sociologists of the twentieth century, does give patients explicit responsibilities. In his 1951 theoretical treatise, *The Social System*, Parsons described "the sick role" as a culturally established set of assumptions governing illness [27].

1. The sick role exempts the patient from normal social responsibilities, typically authorized by the physician.
2. The sick person cannot be expected to get well simply through an act of will.
3. Being sick is socially undesirable, so the sick person has an obligation to get well.
4. The sick person must seek technically competent help and cooperate with the physician.

This is equivalent to an implicit bargain, in which society gives a sick person the right to shirk some ordinary duties, but in return for accepting the authority of physicians in an attempt to recover from the illness and resume customary obligations. Parsons even discussed the problematic situation if a patient doubts what one doctor tells him, suggesting the patient does not have the right to shop around in search of advice. We need not accept the details of the Parsonian analysis, yet the fundamental idea has merit, that in conventional society a person acquires temporary rights and duties during illness, and violating either has ethical implications.

In the case of communicable diseases, the patient probably takes on an obligation to avoid infecting other people, which entails giving complete and accurate information to the doctor and following the doctor's instructions to minimize the chance of contagion. In the context of an epidemic or fear of one, modern information and communications technologies are used to track a potentially infected person's movements in a very public manner, facilitating quarantines and other measures that may harm the individual's reputation whatever the public health benefits may be [28]. More debatable, but following similar logic, the patient may also have an obligation to provide information that would allow medical researchers to develop cures for other patients, even those who have no connection to the original patient.

Early in the history of medical science, the cause and effective treatment of a disease was often discovered through research on relatively small numbers of patients, and the limited environments surrounding them. Today, many of the disorders that still resist understanding and cure may require research studies using very large numbers of patients, even large factions of the whole population. If so, citizens may have the responsibility to share much personal data with their doctors and with health researchers, even without the motivation provided by an acute disease.

17.3 INFORMATION DEVIANCE

Psychiatry, the focus of innumerable ethical and intellectual debates over the past two centuries, is poised to undergo yet another controversial episode with the application of several information technologies. Indeed, the history of psychiatry is like a roller-coaster ride, in which a period filled with optimism is followed by one of pessimism, and each wave of optimism is expressed through a new theory and treatment modality promoted by medical entrepreneurs whose motivations mixed self-aggrandizement with genuine concern for the well-being of their patients. The

very fact that psychiatry is set somewhat apart from other medical specialties highlights some issues that may be less visible in other specialties, yet are present to some degree in all. Perhaps most relevant in the context of ethics, psychiatry is the medical approach best prepared to understand why some people may tend to violate ethical norms. Indeed, psychiatric disorders might be defined as specific failures of information processing in the human brain, systematic errors that could in principle be modeled through artificial intelligence methods [29].

An early example involving historically significant information technologies was the period of building insane asylums in the United States in the mid-nineteenth century. This was an optimistic era for psychiatry, as the *nervous breakdown* theory asserted that unusual stress, especially when experienced by someone with a weak nervous system, could drive a person into a temporarily unhealthy mental condition. Initially, the mental hospitals were conceptualized literally as *asylums*, where the patients would be insulated from stress so that their nervous systems could repair themselves and resume a healthy mental equilibrium. Information technology manifestations of this optimistic ideology were the US censuses of 1850 and 1860, which asked enumerators to record the distinctive cause of each patient's insanity, typically an emotionally stressful life circumstance [30].

In retrospect, the nervous breakdown theory facilitated institutionalization of many people who were not helped by incarceration, and often may have suffered from it. Theories have ethical implications, and all theories in psychiatry raise issues concerning the freedom of the patient to make his or her own treatment decisions [31]. This is not to say that scientific theories are merely ideologies. Some theories about some psychiatric disorders may be factually correct, and among those correct theories some may suggest effective treatments. Rather, the point is that issues of fairness, power, and compassion are implicated in any psychiatric diagnosis and, in the modern age, are shaped by the information system associated with each approach.

The mid-twentieth century saw a very different wave of optimism, associated with the Psychoanalysis Movement. Sigmund Freud did not really claim that all mental disorders were caused by childhood traumas, and could be cured by spending 2 h a week for 2 years on his couch. Yet the implication of his social movement was that this was often the case, and that the treatment was a form of personalized scientific research in which bringing subconscious information to consciousness was itself curative. Given the fact that psychoanalytic thinking is currently in disfavor among academic psychologists, we can wonder about the ethics of a movement that gave lucrative employment to the psychoanalysts [32]. But a more striking ethical issue is the well-documented fact that client-centered psychoanalytic treatment was available for middle-class patients, while working-class patients were subjected to rigid treatments that sought to control their deviant behavior rather than help them realize their inner desires [33].

Originally heavily influenced by psychoanalysis, the American Psychiatric Association developed its *Diagnostic and Statistical Manual* (*DSM*) in 1952, but a 1980 revision reduced this influence while asserting a disease conception of mental deviance [34]. The most visible area of dispute was whether to count homosexuality as a disorder or an example of human diversity, and branding it psychiatric tended to

disparage homosexuals and devalue their contributions to society. More generally, psychiatric orthodoxy contradicts the frequent observation of anthropologists that standards of normality vary across cultures, even such that qualities judged normal or even praiseworthy in our society are condemned in other societies [35, 36]. Modern society is very different from the hunter–gatherer form of life in which humanity originally evolved, and thus the culture within which psychiatrists frame their diagnoses may itself be pathological. At the extreme, psychiatry is used as a tool of social oppression by the elite, and in modern society the information technologies can serve like the whips or shackles of oppressors [37, 38].

Complicating the situation was the rise of an Anti-Psychiatry Movement in the 1960s, represented by *The Myth of Mental Illness* by Thomas Szasz, and many other publications [39]. A number of social scientific studies, varying in their rigor but many worthy of some credence, were published about the unreliability of psychiatric diagnosis and treatment [40, 41]. In sociology, labeling theory argued that a psychiatric diagnosis could be a harmful self-fulfilling prophecy [42, 43]. *Medicalization* of deviant behavior may serve no good purpose [44].

For years after that disruptive period, the situation seemed to have stabilized, with medication providing apparently effective treatment for some sufferers, the Anti-Psychiatry Movement retreating from prominence, and research studies finding some merit in each of the competing theories [45]. But now two information technology-related developments foreshadow a period of uncertainty and debate: (i) The National Institute of Mental Health (NIMH) has severely criticized traditional diagnostic methods and is promoting a new approach that will require extensive genetic and behavioral data on the general population. (ii) A number of scientists and engineers are developing a range of information technologies intended as therapeutic tools and in some cases as permanent assistive technologies to compensate for mental disabilities. Again psychiatry offers clear examples of potential ethical problems that may also arise in other fields.

17.4 THE CURRENT FRENZY

In an April 29, 2013, blogpost titled "Transforming Diagnosis," the head of the NIMH, Tom Insel, severely critiqued the fifth edition of the *DSM*, which was about to be published [46]. While the *DSM* is designed to standardize diagnostic categories, it was largely based on observation of symptoms, rather than on any definitive tests. Insel reported that other fields of medicine had moved away from merely treating symptoms, to the use of objective tests to identify the causes of the specific disease and design a treatment accordingly. He reported: "NIMH has launched the Research Domain Criteria (RDoC) project to transform diagnosis by incorporating genetics, imaging, cognitive science, and other levels of information to lay the foundation for a new classification system." This effort was based on four connected assumptions:

1. A diagnostic approach based on the biology as well as the symptoms must not be constrained by the current *DSM* categories.

2. Mental disorders are biological disorders involving brain circuits that implicate specific domains of cognition, emotion, or behavior.
3. Each level of analysis needs to be understood across a dimension of function.
4. Mapping the cognitive, circuit, and genetic aspects of mental disorders will yield new and better targets for treatment.

The second point, that mental disorders are biological in nature, may seem controversial, yet it makes more sense if we realize that the NIMH would probably exclude from any future revised *DSM* many of the categories that would in prior decades be called *neuroses* or *character disorders*, and it would avoid labeling all human problems under the term *psychiatric*. If the RDoC is fully successful, we might wind up with fewer cases being considered medical issues to be treated by psychiatry, but much better understanding and treatment within that narrower scope.

Insel's blog cites three supportive journal publications, most prominently a review essay in *Nature* by David Adam, "On the Spectrum," that begins by reporting research results contradicting the century-old distinction between schizophrenia and bipolar (manic-depressive) disorder [47]. Other traditional categories seem also to blend into each other, and a popular term like "depression" may not really have one clear meaning. If it is time to junk the old categories, then a number of possibilities open up.

Perhaps some pathologies exist along a spectrum of degrees, as we already now speak of *autism spectrum disorder*, and for decades the ordinary Big Five dimensions of personality identified by psychologists have been conceptualized as matters of degree, one of them being *neurotic* [48]. A given case may, then, be the confluence of several such dimensions on which extreme values are harmful, for example, *neurotic* plus *introverted*. Already back in the late 1930s, it was known that schizophrenia and depression had different distributions across society, with schizophrenia concentrated in the lower class, but it may turn out that the psychiatric distinction made between schizophrenia and depression partly represented social-class prejudices in the minds of psychiatrists [49].

Very recent work in behavioral genetics offers the hope that a new objective set of disease categories may be discovered, through techniques that can be adapted for unambiguous diagnosis of particular cases. It may be that certain genetic conditions predispose a person more or less to exhibit specified psychiatric symptoms, blurring but not totally erasing old distinctions like schizophrenia versus bipolar [50]. If this genetic revolution results in a clear distinction between mental deviance that has powerful biological causes and mental deviance that lacks a biological basis, one result may be a radical shift in how society labels those who in the past would be lumped into a broad category called *mentally ill* or *insane* [51]. Those for whom genetic tests indicate a biological basis for their disorder will be forgiven and treated leniently by the courts and by their neighbors. Those who seem to lack a biological cause could be blamed for their misbehavior and punished, whether through formal imprisonment or informal ostracism from society.

Substantial success in RDoC could have many implications for treatment. Most obviously, the current weak effectiveness of a medication may result in great measure from administering it to populations of patients who are poorly diagnosed. For

example, people suffering a wide range of biological and nonbiological problems are labeled with the vague term "depression," while only subsets of that group have conditions that will respond well to anti-depressive medications. Proper functioning of the human brain has structural and chemical aspects, and some structural problems may not respond well to medications. The interaction between neuronal structure and brain chemistry is complex, so very sophisticated research will be required to unravel its mysteries. There is a reason to hope, however, that some disorders with different ultimate causes may affect behavior through a common pathway, such as influencing what kinds of experience the individual finds rewarding, and thus might benefit from the same treatment [52].

Some disorders may turn out to be predominantly structural and resistant to treatment with pharmaceuticals. For example, in recent years researchers in animal as well as human behavior have postulated the existence of *mirror neurons*, a group of brain cells with the specific function of modeling the behavior of other individuals. They are postulated to be the basis for human empathy, and thus for ethics [53]. How correct this theory really is, especially how distinct the mirror neurons are from other neurons, remains in doubt. But assume for the sake of example that a distinct kind of neurons do function this way. If those neurons are defective or missing in an individual, conceivably causing autism or sociopathy, a medicinal cure may be impossible.

The NIMH-led effort to place psychiatry on a more solid scientific basis may have very substantial societal consequences. A simple triage analysis offers some insights. Suppose the people traditionally considered "mentally ill" are divided into three groups: (i) those with biological disorders that respond to conventional medical treatment, (ii) those with biological disorders that do not respond, and (iii) those whose problems are not biological in nature.

The first category, people who can be helped by medications or other traditional treatments, and who possess a biological problem, fit the sick role defined by Parsons. In a modern liberal democracy with a public healthcare system, they deserve medical treatment and have an obligation to accept it. If the treatment falls short of perfectly compensating for the biological disorder, society has some obligation to make allowances for the individual's disability, as it does for cases of "mental retardation" and for disabilities that are not mental. Information technology plays a key role in the fundamental science, the diagnosis of the individual case, and in monitoring the person's condition over time so that treatment may be adjusted.

The second category may not be entirely hopeless, but raises serious issues. The person may not be competent to play the sick role, and the doctors may have little help to offer. Society's institutions are poorly designed to accommodate extreme biological divergence from the norm, in providing appropriate services for permanently mentally disabled individuals, in schools, housing, and indeed in the criminal justice system. A thoroughly psychiatric conception of human deviance would assert that no criminal deserves punishment, because the cause of deviance is not free will but a lack of mirror neurons, excessive aggressive chemistry, or the like. In addition to the roles played in the first category, future information technology may be implicated in untreatable biological cases as a virtual substitute for mental hospitals and penitentiaries. Already, locked ankle bracelets can monitor the locations of paroled

criminals, and more sophisticated versions of the technology could monitor the behavior of uninstitutionalized mental patients, even against their will [54].

The third category, people whose deviance is not conclusively caused by a well-defined biological defect, raises a variety of perplexing questions. A range of treatments may prove to have no basis in medical science, but to be the secular equivalent of religion. In its origins, psychoanalysis was in some respects a descendant of the *mesmerism* cult, and more than one knowledgeable observer has linked it to specific other mystical religious traditions [55, 56]. Jacob Moreno, a key founder of group therapy, always wanted to found a new religion [57]. Scientology has been the target of much criticism from "respectable authorities," yet may be no less scientific than psychoanalysis, despite defining itself as a religion [58]. It uses an electronic device called the *e-meter* in many of its therapy sessions, an information technology device measuring galvanic skin response and thus related to a lie detector. If the effort of the NIMH casts many mild disorders adrift from medicine, then a whole host of new therapeutic movements may arise, some scientific and some pseudoscientific, and many of the newest therapy fads may be based on information technology.

17.5 GENETIC INFORMATICS

Information technology plays a role in most versions of at least four different stages in the development and use of genetically oriented medical treatments: (i) gene sequencing itself, (ii) research on the medical implications of an individual's genetic makeup, (iii) research on effective treatments for genetically-related disorders, and (iv) management of genetic data about individual humans. The ethical issues are more obvious in stages (iii) and (iv), but the earlier stages must be accomplished first, and thus are indirectly implicated in all the ethical issues.

One of the best sources of questions and insights about the ethical challenges in this area is the report of a conference that took place way back in 1975. Titled *Ethical and Scientific Issues Posed by Human Uses of Molecular Genetics*, its report went far beyond the famous Asilomar conference of that year, which focused on the problems involved in genetic engineering research on micro-organisms, especially issues of laboratory containment of infectious diseases that may be rendered more potent by the research [59]. At that point in time, aggressive genetic engineering dominated both scientific and public discussions of ethical issues, thus intensifying the passions of the debaters, but a range of sophisticated questions were also raised. The introductory chapter of the *Human Uses* report was a thoughtful essay by Daniel Callahan from the Institute of Society, Ethics and the Life Sciences [60]. He recognized that the consequences of scientific research can never be fully foreseen, and then suggested the following four points that place research in the context of general moral principles:

1. Individuals and groups are ordinarily responsible only for the consequences of those actions that are voluntary and intentional on their part. However, they may also be held responsible for the unintended consequences of their actions if, through negligence, they failed to take into account such consequences.

2. Individuals and groups cannot be held responsible for those actions the consequences of which are totally unknown. However, if they voluntarily undertake such acts, they may be held responsible for the consequences unless there were serious reasons for undertaking the action in the first place.
3. When others may be affected by our actions, they ordinarily have a right to demand that their wishes and values be respected.
4. Individual scientists and scientific groups are subject to the same norms of ethical responsibility as those of all other individuals and groups in society.

Later, Callahan considered two other points that were discussed at the conference and still have relevance today. First, he argued that concern for ethics should not be limited to the applied sciences that intended to have practical applications, but to "pure science" as well, because discoveries even in the most abstract fields may eventually lead to significant social consequences. This logically invokes the second of his numbered principles, that any pure science project requires the researchers to articulate serious justifications, which often may be difficult to provide.

Second, Callahan argued that in any debate over the ethics of a particular research project the *burden of proof* should be on those who oppose the research, not those who wish to perform or support it. He based this claim upon the observation that currently the society supported both pure and applied science, implying that the ethics of science must be conducted within a framework defined by the values and beliefs of the general society. Recently, especially in Europe, leaders concerned with the ethics of scientific research have advocated the opposite approach, called the *precautionary principle*: Research should not be done if it plausibly will lead to more harm than good, even in the absence of certainty [61].

The 1975 *Human Uses* conference emphasized issues around *genetic engineering*, or to use a more bland term *gene therapy*. Over the subsequent decades, genetic engineering of human beings has been off the radar of scientists and philosophers alike, and the chief disagreements concerned genetic engineering of the plants and animals used by humans for food. One recent study suggested that nations differ in their tolerance for risk, and this shapes specific policy decisions about genetically modified organisms (GMOs) [62]. One reason gene therapy for humans has lurked in the background, rather than being hotly debated, is that techniques to accomplish the more ethically questionable applications were lacking, most especially information about the genetic makeup of particular individuals and the health consequence of various genetic structures. The new genetic research methods, such as those promoted by the NIMH, will trigger debate, sooner or later, and these methods make heavy use of information technologies [63].

For example, research to identify the genetic correlates of traditionally distinguished psychiatric disorders requires data on very large numbers of research subjects. In one of the studies considering the validity of the traditional psychiatric diagnostic categories, done by the Cross-Disorder Group of the Psychiatric Genomics Consortium, genetic data on more than 60,000 people were required [64]. In the case of mental disorders especially, the privacy of the individual is crucially important.

If it were known that an asymptomatic young person had a genetic structure predictive of psychosis, that individual might suffer employment difficulties, social ostracism, and a variety of other disadvantages connected to his or her socioeconomic situation.

In 2014, US federal government agencies were developing a new *common rule* for the protection of human subjects, a set of principles heavily influenced by the contingencies of medical research and largely motivated by concerns about the radical implications of genetic data. One contribution to this complex debate was a workshop report by social and behavioral scientists, expressing a range of doubts concerning whether definitive rules developed in the medical area could be applied without modification in their own areas [65]. Thus, the ethics of medical informatics is important not only in terms of patient welfare, but as a standard to which other realms of science may conform.

The rapid development of gene sequencing technologies, and the widespread hopes for great future advances in this area, has reminded social critics of the disastrous eugenics movement of around a century ago. The term *eugenics* was coined by Sir Francis Galton, admittedly one of the greatest scientists of his era, and defined in a footnote to his 1883 book: *Human Faculty and Its Development*:

> That is, with questions bearing on what is termed in Greek, *eugenes*, namely, good in stock, hereditarily endowed with noble qualities. This, and the allied words, *eugeneia*, etc., are equally applicable to men, brutes, and plants. We greatly want a brief word to express the science of improving stock, which is by no means confined to questions of judicious mating, but which, especially in the case of man, takes cognisance of all influences that tend in however remote a degree to give to the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had. The word *eugenics* would sufficiently express the idea; it is at least a neater word and a more generalized one than *viviculture*, which I once ventured to use [66].

Galton also defined *eugenics* as *the cultivation of race*, probably intending *race* to mean inheritance very generally, rather than using the word as commonly is done today. His research built upon Darwin's theory of evolution by natural selection, but may also have been influenced by Europe's feudal traditions, as *noble qualities* evoke images of the *good breeding* that aristocrats believed they possessed. In defining *eugenics*, as in much of his work in statistics and other fields, Galton did not strictly distinguish science from technology. In principle, information technology could improve our understanding of human diversity, without requiring us to take any action based on that knowledge. The Eugenics Movement became rather influential within the medical community early in the twentieth century, connected with concerns that modern society was allowing "biologically inferior" people to reproduce at high rates, and justifying even forced sterilization.

Already by 1916, leading anthropologist Franz Boas had attacked the Eugenics Movement for wrongly assuming that many differences between human groups were genetic, when in fact they were the result of environmental differences, for example,

children raised in impoverished households [67]. Yet today technologies exist that might be able to identify which traits were wholly biological in nature. In *The Journal of Medical Ethics*, Daniel Wikler has argued thus:

> We should avoid an unthinking rejection of every eugenic thought or value. The fact that eugenicists were in favour of a particular measure or goal is not in itself sufficient reason to oppose it. We need a good analysis of which eugenic aims were wrong-headed, and why. We might judge that some of the questions to which eugenicists proposed answers ought not be ignored, and indeed that they are now given too little attention, in part because of their eugenic associations [68].

Writing in the same journal, Stephen Wilkinson analyzed the often inflammatory rhetoric used by both sides in any eugenics debate, seeking ways to preserve individual autonomy in making reproductive decisions [69]. The website of the National Human Genome Research Institute that seeks to inform the general public about Huntington's disease concludes with this grim observation: "Prospective parents consider prenatal testing when one parent has been diagnosed with Huntington's disease or has been found to carry the gene. Prenatal testing can show whether the child will inherit the defective gene. To test the fetus, DNA is extracted from fetal cells via CVS (chorionic villi sampling) or amniocentesis. If the fetus tests positive, parents can make decisions about whether to terminate the pregnancy" [70].

It is all very well to say that individuals should be free to make their own choices, including either aborting or bringing to full term a fetus that has been genetically tested and found to have a condition that would render it incapable of self-supporting citizenship as an adult. But in nations where the government takes over much of the responsibility for funding medical treatment, Parsons would argue the obligations of the patient expand. Consider the following four claims that might be made in a debate about ethical implications of genetic information, not because they are all reasonable but because they highlight real issues:

1. Because genetic testing may eventually lead to eugenics, in which the slippery slope of public health care leads to imposition of inhumane policies, genetic testing should be forbidden.
2. If a pregnant woman knows, on the basis of genetic evidence, that her child would be dependent upon government support throughout its life, she should be required to abort it, and if she refuses, the child will not be eligible for publically funded medical treatment or other benefits.
3. Because exposure of genetic data can lead to severe social, economic, and psychological consequences, anyone who makes such data about another person public is guilty of a felony deserving long imprisonment.
4. Because public disclosure of genetic information about any one individual could be harmful to blood relatives of the individual, no person should be allowed to make public their own genetic data, or indeed to have their own genome sequenced because doing so violates the privacy of parents.

The first two points, expressing two competing concerns about a resurgence of the eugenics movement, are not fanciful, and a recent article in the *American Sociological Review* raised this specter [71]. One of the earliest research successes of the new genetics was the discovery of the genetic factor in sickle cell anemia among people of African descent. Duana Fullwiley's research indicated that this concept influenced developments in the culture of Senegal that could significantly affect interpersonal relations and posed difficult policy issues for the government of that nation [72]. She also explored how medical genetics could stimulate racism, with a specific focus on research concerning rates of asthma in Latino populations [73].

Much of the recently publicized progress in genetic research relevant for medicine has focused on single nucleotide polymorphisms (SNPs) often pronounced *snips*. Calling them *snips* is a nice metaphor, because they are the simplest and smallest kind of genetic variation. In principle, some rare SNPs may have a very significant effect on the biological development of an individual who possesses one, rendering genetic diagnosis of a disorder relatively straightforward, even if they only increase the probabilities of the disorder [74].

However, many kinds of human variation are the results of various genes interacting as a complex system, blurring the lines between categories, and opening possibilities for external factors to play decisive roles. For example, for many people the symptoms of diabetes are affected by diet and exercise, as well as by the person's genetic makeup—the so-called type II diabetes. It has long been recognized in the area of psychiatry, the locus of the NIMH's aggressive genetic research, that biology and environment may interact in complex ways to produce abnormal behavior [75, 76]. Thus, in many areas of medicine some genetic diagnoses will not be cognitively simple, assigning people to strict categories, but assessing various degrees of risk factors. This situation makes it difficult to give patients authoritative advice, and increases the importance of medical informatics distributed widely via information technology.

17.6 UBIQUITOUS INFORMATION TECHNOLOGY

Consider the example of a member of the general public who is already being medicated for a heart condition, whose doctor has just prescribed that niaspan should be added to the mix. This person opens a Web browser and enters "niaspan" into Google. The first hit is www.niaspan.com, which explains niaspan is intended to raise "good cholesterol" in the bloodstream and reduce "bad cholesterol." Then the site immediately addresses "safety considerations," and at some length lists potential problems, including symptoms of serious side effects that should be brought to the attention of a physician. Clearly, this is entirely ethical, in the context of conventional medical practice, leaving the ultimate decision about medication to a discussion between the patient and a doctor. But suppose the person does not trust the physician, or has been influenced to distrust the entire medical profession, either on the basis of bad past personal experience, or distressed by public debates such as the intense arguments about government-provided health care in the United States.

So, the person looks further down the hits returned by Google, and finds a thoughtful news article by Harlan Krumholz on the www.forbes.com website, "Five Lessons from Niaspan's Disappointing Study" [77]. It discusses the difficulty of carrying out conclusive scientific studies of the effectiveness of medications, and reports that a recent study, supported both by the pharmaceutical company that manufactures niaspan and by the National Institutes of Health, was halted before completion because niaspan did not seem to provide any benefits greater than those offered by a placebo, and because the side effects made it potentially harmful for some patients. Uncertain whether to trust *Forbes*, given its presumed political leanings, the person continues to search, finding a similar article by Gardner Harris in *The New York Times*, "Study Questions Treatment Used in Heart Disease" [78]. Being a savvy Internet user, and seeing that both articles date from 2011, the person now googles "niaspan 2011 site:nih.gov" and finds a press release from the National Institutes of Health, "NIH stops clinical trial on combination cholesterol treatment: Lack of efficacy in reducing cardiovascular events prompts decision" [79].

We cannot here settle the issue of whether niaspan might be valuable for many patients, despite the dismal results of this particular clinical trial, but will use it as an example of the perplexities that face ordinary citizens trying to make medical decisions in the Internet era. Careful reading indicates that the patients in the clinical trial had a particular range of heart-related symptoms, and both those getting niaspan and those getting the placebo were also taking other medications that might have masked the benefit of niaspan. A more general insight is that many medications may improve the results of medical tests, such as the balance of cholesterol in the blood, without actually having any health benefit. Are abnormal cholesterol levels in the blood a primary cause of circulatory trouble, or merely an indicator of some more fundamental cause? Could the answer be "both, but under different conditions for specific patients?" Questions like this raise deep philosophical issues about causes and effects in complex systems. When causal relationships are poorly understood, ethical rules will be poorly framed.

Reading the online articles, the person might discover that niaspan is a version of a common vitamin, niacin, in an extended-release form so that a high dose can be given without shocking the system, and be absorbed gradually. So, what is the difference between medications and nutritional supplements? The medical profession plays an important role in classifying products to distinguish: (i) nutritional supplements, (ii) over-the-counter medicines, (iii) prescription medicines, or (iv) addictive drugs. But governments make the ultimate decisions about what the categories will be and which products will be assigned to each. Today, information technology is playing a role in classifying not only medications but also every noun, verb, or adjective used in medical discourse. Already a decade ago, the computer-based Unified Medical Language System developed by the National Library of Medicine contained 2.5 million terms, 900,551 concepts, and more than 12 million relations between these concepts [80]. While admirable, efforts to reduce such huge lexicons to manageable size without introducing grievous errors have achieved only modest success.

A radical approach has emerged in recent years, both naturally among the general public and in carefully designed academic research studies, in which ordinary people

advise each other on health-related issues, rather than relying entirely upon their doctors. Logically enough, much of the online communication activity has concerned chronic problems that resist cure but demand management. Studies have explored the opportunities for creating software or other systems that could guide people in *collaborative tagging*, for example, in which an online network of people might label various foods for their nutritional value [81].

Mobile devices are already playing valuable roles in managing health problems, such as diabetes, for which data must constantly be collected and patients must be guided in dealing effectively with their own conditions [82, 83]. Already today, many people use special software on mobile devices to monitor and improve their health-related behavior, and much of the best research builds on well-established theories of behavioral change [84]. Some of the influential recent studies incorporate the technology in the community and culture to which the individual belongs, thus constituting a new form of social computing [85].

One of the more active research areas has focused on autism spectrum disorders, some of it based in the field of affective computing, the theories and methods that concern ways by which computer systems could understand, emulate, or at least adapt to human emotional expressions [86, 87]. The long-term results of such research could include (i) mobile devices that assist an autism spectrum person in their day-to-day interaction with other people, (ii) training systems to teach children suffering this disorder how to read and respond properly to the emotions of other people, and (iii) research methods for the study of human emotional communications more generally [88, 89]. Depression and epilepsy are among the other disorders that have been the focus of research developing mobile applications [90, 91].

Today, many people are *cyborgs*, in the sense that they are combinations of biological organisms and information technology: They have heart pacemakers. Tomorrow, many people may have brain pacemakers, whether implanted in their otherwise useless sinuses or attached to their bodies externally like hearing aids. Some of these machines may merely advise autism spectrum disorder sufferers in their social relations, while others will employ electrical stimulation of the brain to combat depression. There will be serious ethical issues about the extent to which people are really free to decide whether or not to become cyborgs, given the social pressures to conform.

For applications such as providing emotional support for elderly people, researchers are developing robots that could be "always-on relational agents," and a software application called MemoryLane helps people reflect about their pasts, a short step away from robot psychoanalysis [92, 93]. To what extent does the proliferation of information technologies erode employment opportunities for human beings, replacing health and service personnel with machines? [94, 95].

Some research from the value-sensitive design perspective has used social science methods to explore how people might judge the ethics and efficacy of a range of mobile device applications through which parents could monitor the activities of their children [96]. This approach frames technology design in terms of the goals which users want it to achieve, developing a design philosophy for the type of application, and doing empirical research on the impact the technology actually has on users [97].

Thus it gives the role of ethics judge to the users, rather than to lawyers, philosophers, or political leaders. Yet nonscientists are influenced by popular fads, fallacies, and corporate advertising. Should humans give over control of their minds to information technologies, no matter how well-intentioned their developers may be?

17.7 STASIS VERSUS PROGRESS

The primary goal of medical science is to improve prevention and treatment, and across all of the sciences the rhetoric of progress raises hopes for beneficial technological applications. Yet in areas such as the space program and nuclear energy, progress seems to have stalled despite vigorous research [98]. Some pundits have suggested we are nearing the end of science, although it seems plausible that one science may reach its limits without others doing the same [99]. The biological and information sciences seem to be immune to stasis, because their subject matter is so complex we cannot believe that all questions will be answered soon. Yet the complexities we have already identified can be cause for worry that medical science is approaching its practical limits. Indeed, while medicine contributed to the tremendous increase in the life span during the twentieth century, economic growth and improved public health based on economic growth did as well, and demographers tend to doubt that improvement can be sustained during the current century [100, 101].

A philosophical argument against endless progress can be based on the principles of biological evolution itself. Every living species occupies an ecological niche, adapted to survive within a particular set of physical constraints. Major developments in evolution tend to produce adaptive radiation, as one species begins to exploit a new environment, thus coming under pressures that cause rapid change and even diversification into multiple species. Perhaps triggered by environmental changes, humanity evolved in East Africa, proverbially "climbed down from the trees," and developed technologies that themselves transformed relations to the environment, thus feeding back to cause even more evolution. However, across the vast history of biological evolution, such periods of rapid change tend to be followed by more static periods, when the life-forms fully adapt to their new ecological niche. The proverbial example is the fact that birds can't fly to the moon. Their evolving wings allowed them to soar above the trees their ancestors inhabited, but that did not enable them to fly beyond the limits of Earth's atmosphere. Similarly, the metaphoric "wings" of technology may exploit but not exceed the limits of that same planet [102].

Vannevar Bush called science the *endless frontier*, asserting that human progress could continue indefinitely if science received vigorous support [103]. Yet the most prominent theory of frontiers, enunciated over a century ago by Frederick Jackson Turner, suggested that eventually every frontier would close, with profound implications for human society, notably a severe weakening in the factors that support democracy [104]. Applied to technological innovation, the closing of the frontier could force development of ever-more harmful technologies. Unwilling to admit that progress had ended, societal leaders would support ever more dubious projects, just

to keep their dreams of power alive. Thus innovations of dubious value may become societal placebos. For some patients, antidepressants may already function as *active placebos*: Their side effects tell the patient that a powerful treatment is in progress, which reduces anxiety and increases hope. Information technology can do the same.

It is hard to think of a more ethically problematic phenomenon than the placebo effect. A lie provides real benefit. One of the most insightful publications is an obscure report by psychotherapist John Snell about his use of hypnosis with patients who believed they were the victims of supernatural hexes [105]. Psychoanalysis included hypnosis among its original treatment methods, because it was in part derived from the mystical pseudoscience of Franz Anton Mesmer, and *mesmerism* is a synonym for *hypnosis* [106]. A patient who subscribed to the folk belief in hexes could interpret hypnosis as exorcism, experience a powerful placebo effect, and become cured. But how ethical is it to deceive the patient, even in service to the patient's needs? If medical science reaches its limits, how ethical will it be to employ deception to improve treatment to at least some modest degree?

Another way to conceptualize the future limits of medicine is to notice a quality that the antique nervous breakdown theory shared with the germ theory of disease. The patient was conceptualized as fundamentally normal, perhaps weaker than others in some respects, but still healthy. Then, the patient was attacked by a disease from external sources: germs swimming in the drinking water, unusual stress damaging the nervous system, or a horse that bucks and breaks the rider's arm. A doctor defeats the attacker or repairs the damage.

Quite different is conceptualization of a genetic or otherwise inherent problem as a disability, but redefining *disability* as *disease* with the implication that a cure must be provided. A competing perspective is that *disability* should be reconceptualized neutrally as *diversity*, with the implication that society and its technologies should equally benefit the full range of human beings, rather than stigmatizing and seeking to transform unusual individuals [107].

Several critics of advanced assistive technologies view them as unfair or dangerous enhancements of human nature, unfair because not all people can afford them, and dangerous not merely because of specific risks but because they violate fundamental human values [108]. The social movement advocating systems to enhance ordinary human abilities through the use of new information technologies call this development *transhumanism* [109]. Opponents accuse this radical attempt to achieve a *posthuman* future of endangering humanity rather than transcending it [110, 111]. Many classic science fiction novels imagine that a small number of enhanced humans might find themselves the victim of persecution campaigns by so-called normal people [112, 113].

More likely, a new competition equivalent to a military arms race between nations might be triggered between parents, who begin merely giving their children assistive technologies to overcome disabilities, then start forcing normal children to be fitted with assistive technologies to compete with their increasingly augmented classmates. Long ago, sociologist Michael Young argued that any form of meritocracy, in which society was run by the best and the brightest, would stimulate extreme political resentment in the general population; but if that meritocracy were achieved through artificial technologies, the result could be bloody revolution [114].

17.8 PROBLEMATIC ETHICS

Now that we have identified some very perplexing ethical issues involving information technology either directly or indirectly, with concrete examples, we can consider more abstractly what the philosophy of ethics might say about them. As illustrated by the fact that people still read the essays by Plato and Aristotle, more than 2000 years after the deaths of these ancient Greeks, philosophy is deeply rooted in traditional human cultures. It could, however, be described as the most radical of the humanities, providing a literary critique of conventional thinking. Thus, great philosophy exhibits a tension between conventionality and radicalism.

Some philosophy does look like science, for example, the discipline of symbolic logic, which is closely related to computer science [115]. However, some of the most influential philosophies of science contradict common views among scientists. Karl Popper argued that while scientific theories can be proven false, they can never be proven true, without implying by this that science was inherently false [116]. Especially relevant to the topic of this chapter, G. E. Moore argued that ethics cannot be based entirely on empirical facts, and may in some fundamental sense be subjective [117]. Moore sometimes referred to ethical principles as the things we desire to desire, suggesting some kind of infinite regress that would never reach solid ground.

Talcott Parsons conceptualized ethics in terms of the *values* enshrined by a particular culture as those things that all members should desire, in order for the institutions of society to function well, with the optimistic view that human progress could bring the entire species to a consensus on what those values should be. In the decades after Parsons wrote, many social and behavioral scientists expressed doubts about the concept of values, for example, arguing that human action was seldom guided by over-arching principles of any kind, let alone ethical ones, but by immediate contingencies of behavioral reinforcement in the environment [118]. From that anti-Parsonian perspective, values were mystical hopes more appropriate for theologians than scientists; but in recent years, the concept has had something of a revival within sociology [119].

Conventional philosophy of ethics often started from the premise that some irreducible set of ethical principles must logically be followed. These were the *categorical imperatives* of Immanuel Kant, rules that must be obeyed by all people under all circumstances, in distinction to norms that apply only under specific contingencies of the form: If X do Y [120]. This Kantian assumption influenced many later philosophers, notably John Rawls, in his widely praised 1971 book, *A Theory of Justice* [121]. However, Rawls and this entire tradition came under criticism from social scientists like George Homans, who argued that human beings are not in fact ruled by over-arching ethical principles, but instead seek personal gain, and over time learn to cooperate with other people, but only in order to fulfill their selfish motives [122].

Parsons and Homans disagreed with each other, in rather public debates, yet they both rejected the notion that ethics are objective. Rather, behavioral norms are practical rules that serve to maximize benefits and reduce costs. Where these two social theorists differed was on the question of whose benefit was paramount. Whereas Homans focused on benefit for the individual, Parsons analyzed human behavior from the standpoint of functionality for the society as a whole [123, 124].

Information science enters the picture in the work of Robert Axelrod, whom Homans privately praised with great enthusiasm [125]. Axelrod's research was an example of computer simulation within the tradition called *game theory* [126]. Specifically, Axelrod wrote a simulation program based on the prisoner's dilemma, a class of interaction games in which two individuals make bargains with each other and either fulfill their obligations or betray each other. He then invited colleagues to submit decision rules that the simulated people would follow, deciding whether to keep the bargain or violate it on a particular turn, often on the basis of past experience. While not couched in terms of ethics, these decision rules can be evaluated in terms of how well they support cooperation between the two simulated individuals. Computer simulations, of the form technically called *multiagent system*s and using neural nets to model individual humans embedded in a large social network, have explained the origin even of religious faith and competing sects [127].

However, humans are biological organisms, rather than computers, so much of their programming is genetic. This takes us into the medically relevant territory of sociobiology. In *The Insect Societies*, E. O. Wilson masterfully showed how social insects evolve cooperative behavior, despite lacking complex brains capable of philosophizing [128]. He then applied these ideas to mammals including human beings [129]. The human brain could not evolve to its current complexity without very cooperative social groups, the family and the hunter–gatherer band, because infants are totally helpless until their brains mature physically and learn complex behaviors over a period of years. The concept of *inclusive fitness* refers to characteristics that evolved over the eons through natural selection from random variation, which benefit not so much the individual but other family members who share many of the same genes.

The situation becomes more problematic when evolutionary biology merges with game theory, as in the work of John Maynard Smith [130]. Yes, a big-brain mammalian species like humans benefits from genes that encourage cooperation, but some individual members of the species benefit from being less cooperative. They can free-ride on the cooperation of others, perhaps directly stealing from them. Thus, deviance from ethical standards is in a higher sense "normal." Ethics are *categorical* in a different sense from that used by Kant. Traditionally, in human history, stricter rules applied to interaction with members of one's own social category, than with outsiders. As much as we today decry nationalism and racism, tribalism is an aspect of human nature. "Us versus them" divides people into two categories, those deserving ethical treatment, and those not.

This line of thinking from evolutionary biology, and the prisoner's dilemma as simulated by information technology, brings us back to the practice of medicine. Ethics is a relatively simple issue, when there are only two parties to the interaction: a medical professional and a patient. In order to earn the confidence of the general public, and have dutiful, paying customers, the medical profession establishes an ethical code. While not simple, the situation can be relatively stable. But coherence falls apart when a very large number of other agents, belonging to professional and political groups with very different interests, transform medical treatment into an arena for conflict.

17.9 LEADERSHIP IN SCIENCE AND ENGINEERING ETHICS

If abstract principles cannot rule the day, then social organizations may fill the gap. Well-established scientific associations typically promulgate explicitly stated ethical rules, yet membership in many of them is voluntary, and enforcement of the rules tends to be lax. The status of scientific organizations can be problematic, especially in the social and behavioral sciences, which perhaps ironically are likely to possess special expertise in analyzing ethics.

While they have identical acronyms, the American Psychiatric Association, the American Psychological Association, and the American Psychoanalytic Association are unrelated, the first APA having been founded in 1844, the second in 1892, and the third in 1911, respectively. In recent decades, two new significant alternatives to the American Psychological Association have arisen. Some academic researchers in this APA criticized it for being too applied, given that many other members were clinical or educational psychologists, so in 1988 the American Psychological Society was formed, later renamed the Association for Psychology Science, keeping the APS acronym but stressing its international and scientific ambitions. A challenge to all these APA and APS groups is the new multidisciplinary field called "cognitive science," and the Cognitive Science Society was formed in 1979. One implication of the CSS is the suggestion that all the other societies devoted to the study of the human mind are based on false or at least incomplete scientific principles. While all these associations could perhaps agree on some of the same ethical principles, many researchers in the fields they compete over may belong to none of them.

Computer and information scientists are similarly split across many organizations. Two prominent ones are the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). As their names suggest, these associations began with a focus on developing computer hardware, but their emphasis shifted somewhat to software and human-centered issues over the years, and both have ethical codes. As of mid-2014, the IEEE Code of Conduct primarily listed the ways in which members should obey the laws that apply to all citizens, but included this statement: "We will be respectful of the privacy of others and the protection of their personal information and data" [131].

The ACM Code of Ethics and Professional Conduct is rather complex and philosophical, listing 24 *imperatives*, using Kant's term but somewhat more broadly than he did. The first eight actually might qualify as categorical imperatives, but the ACM calls them *general moral imperatives*:

1. Contribute to society and human well-being
2. Avoid harm to others
3. Be honest and trustworthy
4. Be fair and take action not to discriminate
5. Honor property rights including copyrights and patents
6. Give proper credit for intellectual property
7. Respect the privacy of others
8. Honor confidentiality [132]

How many of those principles might apply to scientists and engineers who develop military applications is unclear, and many other aspects of human life involve using technologies to gain advantage in conflict against other human beings. The fact that privacy and confidentiality are two separate principles hints at an issue for government security agencies that use much advanced information technology, such as the Federal Bureau of Investigation, the Central Intelligence Agency, and the National Security Agency. They may violate people's privacy, by gaining personal information that people do not want to share about themselves, yet uphold the confidentiality principle by not making damaging information public. In at least one highly publicized case concerning the controversial diagnosis of "attention deficit hyperactivity disorder" in children, researchers destroyed data rather than provide them in compliance with court actions [133]. Once data exist in electronic form, it is easy to copy them, so disagreements among the members of a research team dealing with a complex set of demands can lead to loss of confidentiality, even before we consider the possibility of outsiders hacking into the database.

A very different perspective has been developed by the Association of Internet Researchers, suggesting that online research is so complex and radically new that an explicit set of rules cannot be established. In a 2012 document, this organization offered a long list of ethical questions that an Internet researcher should consider, suggesting a "bottom-up" rather than "top-down" approach, in which each individual should develop a personal ethical code:

> At its most fundamental level, we recognize that ethical decision-making interweaves one's fundamental world view (ontology, epistemology, values, etc.), one's academic and political environment (purposes), one's defining disciplinary assumptions, and one's methodological stances. Decision making occurs at many junctures in the cycle of inquiry, including research design, research conduct, and research production and dissemination [134].

The Association report specifically noted that many of the widely held principles of scientific ethics were derived from traditional medical research, in which there was a rather clear definition of what a human subject was, and in which some form of social contract between the human subjects and the researcher existed. In much online research, the people under study are not asked to give informed consent, and in fields as diverse as criminology and political science there may exist an inescapable conflict of interest between the researcher and the people under study.

The most egregious examples connected to criminology do not involve academic researchers, but "criminals" and law enforcement, each apparently prepared to set aside widely held moral principles in order to accomplish practical tasks. A noteworthy harbinger of things to come may be the recently emerged "mugshot industry" in which websites post arrest photographs, which many jurisdictions believe are public information and thus must be shared. Some companies charge people to remove a discrediting mugshot, which apparently is legal in most jurisdictions rather than constituting extortion because formal definitions of that crime limit it to cases in which a person pays to prevent something from being made public, not to remove something that is already public [135]. Many mugshots are of people who were

found innocent after arrest, or involve minor offenses, notably public drunkenness among those related to the person's health. Given the shame some of these pictures may cause, including not merely the ugliness of some portraits but disease-related information associated with them, local jurisdictions must consider changes to the laws that govern distribution of police information about arrested persons. Another recent trend has been the proliferation of websites that allow patients to rate their physicians, which may attract that nasty fraction of the online population called *trolls*, who delight in posting hostile comments, whether or not their judgments are based on objective facts.

A prominent example of the political dimension is the Truthy research done at Indiana University, with funding from the National Science Foundation, which in 2014 came under Congressional scrutiny with accusations in the popular press, either that it was motivated by a particular political agenda rather than scientific values, or that it sought ultimately to give government unusual powers to monitor and oppress its opponents. This was actually only one of many NSF-funded projects developing new methodologies for studying online political communications, several of which employed the readily available Twitter data stream. Some political messages sent via Twitter were automatically generated, rather than having been "tweeted" by people, and a major research question for several parallel projects was how to recognize and understand the dynamics of false information spread through social media. As it happens, several of the publications from Truthy scientists identified Republican rather than Democratic organizations as the source of false information and robot messaging, and this fact may have encouraged Republicans to be especially suspicious of the project [136, 137]. The methodology is likely to be applied in future to online messaging about health issues, whether critiques of government programs, or partisan interpretations of public health issues.

Less conflictual but potentially controversial is the development of computerized information systems to provide advice to people when they are making decisions involving ethical issues. As early as 1987, leaders in the British Medical Association were promoting COnsent to MEdical Treatment (COMET), what was then called an *expert system*, that today might be called a *decision support system*, to advise medical professionals about ethical issues [138]. Efforts along these lines have continued ever since [139, 140]. However, such systems typically make the egregious error of reducing ethics to a set of simplistic rules, based on vague definitions that are difficult to apply in real-world situations, and on assumptions that only some groups in society accept [141].

17.10 CONCLUSION

Philosophical and social scientific analysis of ethical issues seldom leads to clear, simple conclusions; this results not from the failure of academic thinking but from the naturally problematic nature of morals. In the absence of an all-seeing, loving God, no source exists for a perfect moral system, and for reasons that must be left to the reader to decide, the modern political system does not seem to listen to any deity.

Yet, as a practical matter, people must obey the equivalent of a social contract in their interactions with each other, and professions must develop codes of ethics, even if they differ in their details. The revolutions associated with computer technologies are many and profound, offering new hopes for doctors, patients, and all the related professions. As healthcare information technology becomes ubiquitous, its designers and users will need to think deeply about many ethical issues.

REFERENCES

[1] Federal Networking and Information Technology Research and Development Program (2014) Report on privacy research within NITRD. National Coordination Office for NITRD, Arlington, VA. Available at: https://www.nitrd.gov/Pubs/Report_on_Privacy_Research_within_NITRD.pdf, accessed February 22, 2016.

[2] Ohm, P. (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review 57(6), 1701–1777.

[3] Schwartz, B. (1968) The social psychology of privacy. American Journal of Sociology 73(6), 741–752.

[4] Brown, M.F. (2008) Cultural relativism 2.0. Current Anthropology 49(3), 363–383.

[5] Michels, R. (1915) Political Parties. Hearst's International Library, New York.

[6] Gilens, M., Page, B.I. (2014) Testing theories of American politics: Elites, interest groups, and average citizens. Perspectives on Politics 12(3), 564–581.

[7] Obama Administration (2012) Consumer data privacy in a networked world. Available at: https://www.whitehouse.gov/sites/default/files/privacy-final.pdf, accessed October 29, 2014.

[8] Lastowka, G. (2010) Virtual Justice: The New Laws of Online Worlds. Yale University Press, New Haven, CT.

[9] Homans, G.C. (1937) Partible inheritance of villagers' holdings. The Economic History Review 8(1), 48–56.

[10] Homans, G.C. (1941) English Villagers of the Thirteenth Century. Harvard University Press, Cambridge, MA.

[11] Weil V., Snapper, J.W. (eds.) (1989) Owning Scientific and Technical Information: Value and Ethical Issues. Rutgers University Press, New Brunswick, NJ.

[12] Hall, B. (ed.) (2012) No Safe Harbor. United States Pirate Party, Jacksonville, FL.

[13] U.S. Department of Health and Human Services (2014) Public access policy. Department of Health and Human Services, Bethesda, MD. Available at: publicaccess.nih.gov/, accessed October 24, 2014.

[14] Hobbes, T. (1651) Leviathan. Crooke, London.

[15] Rousseau, J.-J. (1795) A Treatise on the Social Compact. Eaton, London.

[16] Nietzsche, F. (1907) Beyond Good and Evil. Macmillan, New York.

[17] Sorokin, P.A. (1937–1941) Social and Cultural Dynamics. American Book Company, New York.

[18] Principles of medical ethics. Available at: www.ama-assn.org/ama/pub/physician-resources/medical-ethics/code-medical-ethics/principles-medical-ethics.page?, accessed November 28, 2014.

[19] Hughes, J.W., Snyder, E.A. (1989) Policy analysis of medical malpractice reforms: What can we learn from claims data? Journal of Business and Economic Statistics 7(4), 423–431.

[20] Currie, J., MacLeod, W.B. (2008) First do no harm? tort reform and birth outcomes. The Quarterly Journal of Economics 123(2), 95–830.

[21] Born, P., Viscusi, W.K., Baker, T. (2009) The effects of tort reform on medical malpractice insurers' ultimate losses. The Journal of Risk and Insurance 76(1), 197–219.

[22] McMillan, R.C. (1995) Responsibility to or for in the physician-patient relationship? Journal of Medical Ethics 21(2), 112–115.

[23] Leveson, N. (1995) Safeware: System Safety and Computers. Addison-Wesley, Boston, MA.

[24] Huff, C.W., Brown, R. (2004) Integrating ethics into a computing curriculum: A case study of the Therac-25. In Akera, A. & Aspray, W. (eds.) Using History to Teach Computer Science and Related Disciplines. Computing Research Association, Washington, DC, pp. 255–277.

[25] Yan, D., Wong, J., Vicini, F. et al. (1997) Adaptive modification of treatment planning to minimize the deleterious effects of treatment setup errors. International Journal of Radiation Oncology, Biology, Physics 38(1), 197–206.

[26] Finholt, T.A., Birnholtz, J.P. (2006) If we build it, will they come? The cultural challenges of cyberinfrastructure development. In Bainbridge, W.S. & Roco, M.C. (eds.) Managing Nano-Bio-Info-Cogno Innovations. Springer, Dordrecht, pp. 89–102.

[27] Parsons, T. (1951) The Social System. Free Press, Glencoe, IL, pp. 436–437.

[28] Gottlieb, S. (2014) Why Ebola quarantines will grow larger—and more troubling, *Forbes*. Available at: www.forbes.com/sites/scottgottlieb/2014/10/30/why-ebola-quarantines-will-grow-larger-and-more-troubling/, accessed October 31, 2014.

[29] Bainbridge, W.S. (1995) Minimum intelligent neural device: A tool for social simulation. Journal of Mathematical Sociology 20, 179–192.

[30] Bainbridge, W.S. (1984) Religious insanity in America: The official nineteenth-century theory. Sociological Analysis 45, 223–240.

[31] Schneider, P.L., Bramstedt, K.A. (2006) When psychiatry and bioethics disagree about patient decision making capacity (DMC). Journal of Medical Ethics 32(2), 90–93.

[32] Bainbridge, W.S. (2011) The psychoanalytic movement. In Bainbridge, W.S. (ed.) Leadership in Science and Technology. Sage, Thousand Oaks, CA, pp. 520–528.

[33] Hollingshead, A., Redlich, F.C. (1958) Social Class and Mental Illness. Wiley, New York.

[34] Rogler, L.H. (1997) Making sense of historical changes in the diagnostic and statistical manual of mental disorders: Five propositions. Journal of Health and Social Behavior 38(1), 9–20.

[35] Benedict, R. (1934) Patterns of Culture. Houghton Mifflin, Boston, MA.

[36] Edgerton, R.B. (1966) Conceptions of psychosis in four East African societies. American Anthropologist 68, 408–424.

[37] Orwell, G. (1949) Nineteen Eighty-Four. Harcourt, Brace, NY.

[38] Medvedev, Z.A., Medvedev, R. (1971) A Question of Madness. Knopf, New York.

[39] Szasz, T. (1961) The Myth of Mental Illness: Foundations of a Theory of Personal Conduct. Hoeber-Harper, New York.

[40] Strauss, A.L., Schatzman, L., Bucher, R. et al. (1964) Psychiatric Ideologies and Institutions. Free Press, New York.

[41] Rosenhan, D. (1973) On being sane in insane places. Science 179 (4070), 50–258.

[42] Scheff, T. (1966) Being Mentally Ill. Aldine, Chicago, IL.

[43] Lemert, E. (1967) Human Deviance, Social Problems, and Social Control. Prentice-Hall, Englewood Cliffs, NJ.

[44] Conrad, P. (1992) Medicalization and social control. Annual Review of Sociology 18, 209–232.

[45] Perry, B.L. (2011) The labeling paradox: Stigma, the sick role, and social networks in mental illness. Journal of Health and Social Behavior 52(4), 460–477.

[46] Insel, T. Transforming Diagnosis. Available at: www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml, accessed November 5, 2014.

[47] Adam, D. (2013) On the spectrum. Nature 496, 416–418.

[48] Wiggins J.S. (ed.) (1996) The Five-Factor Model of Personality: Theoretical Perspectives. Guilford Press, New York.

[49] Faris, R.E.L., Dunham, H.W. (1939) Mental Disorders in Urban Areas: An Ecological Study of Schizophrenia and Other Psychoses. University of Chicago Press, Chicago, IL.

[50] Craddock, N., Owen, M.J. (2010) The Krepelinian dichotomy—going, going … but still not gone. The British Journal of Psychiatry 96, 92–95.

[51] Gove, W.R. (2004) The career of the mentally ill: An integration of psychiatric, labeling/social construction, and lay perspectives. Journal of Health and Social Behavior 45(4), 357–375.

[52] Dichter, G.S., Damiano, C.A., Allen, J.A. (2012) Reward circuitry dysfunction in psychiatric and neurodevelopmental disorders and genetic syndromes: Animal models and clinical findings. Journal of Neurodevelopmental Disorders 4(19), 1–43.

[53] Iacoboni, M., Molnar-Szakacs, I., Gallese, V. et al. (2005) Grasping the intentions of others with one's own mirror neuron system. PLoS Biology 3(3), e79.

[54] Roman, J.K., Liberman, A.M., Fontaine, J. et al. (2012) The Costs and Benefits of Electronic Monitoring for Washington, D.C., Urban Institute, Justice Policy Center, Washington, DC.

[55] Bakan, D. (1958) Sigmund Freud and the Jewish Mystical Tradition. Van Nostrand, Princeton, NJ.

[56] Berkower, L. (1969) The enduring effect of the Jewish tradition upon Freud. American Journal of Psychiatry 125(8), 1067–1073.

[57] Hare, A.P., Hare, J.R. (1996) J. L. Moreno. Sage, London, p. 5.

[58] Lewis, J.R. (ed.) (2009), Scientology. Oxford University Press, New York.

[59] Berg, P., Baltimore, D., Brenner, S. (1975) Summary statement of the Asilomar Conference on Recombinant DNA Molecules. Proceedings of the National Academy of Sciences 2(6), 1981–1984.

[60] Callahan, D. (1975) Ethical responsibility in science in the face of uncertain consequences. In V, M. & Morison, R.S. (eds.) Ethical and Scientific Issues Posed by Human Uses of Molecular Genetics. Ahab, Glencoe, CA, pp 1–12.

[61] von Schomberg, R. (2012) The precautionary principle. In Bainbridge, W.S. (ed.) Leadership in Science and Technology. Sage, Thousand Oaks, CA, pp. 397–405.

[62] Wohlers, A.E. (2010) Regulating genetically modified food: Policy trajectories, political culture, and risk perceptions in the U.S., Canada, and EU. Politics and the Life Sciences 29(2), 17–39.

[63] Green, E.D., Guyer, M.S. (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470 (7333), 204–213.

[64] Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013) Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis The Lancet 81(9875), 1371–1379.

[65] Pool, R. (ed.) (2013) Proposed Revisions to the Common Rule: Perspectives of Social and Behavioral Scientists. National Academies Press, Washington, DC.

[66] Galton, F. (1883) Inquiries into Human Faculty and Its Development. Macmillan and Co., London, pp. 24–25.

[67] Boas, F. (1916) Eugenics. The Scientific Monthly 3(5), 471–478.

[68] Wikler, D. (1999) Can we learn from eugenics? Journal of Medical Ethics 25, 183–194.

[69] Wilkinson, S. (2008) 'Eugenics talk' and the language of bioethics. Journal of Medical Ethics 34(6), 467–471.

[70] National Human Genome Research Institute. Learning about Huntington's disease. Available at: www.genome.gov/10001215, accessed October 31, 2014.

[71] Phelan, J.C., Link, B.G., Feldman, N.M. (2013) The genomic revolution and beliefs about essential racial differences. American Sociological Review 78, 167–191.

[72] Fullwiley, D., (2010) Revaluating genetic causation: Biology, economy, and kinship in Dakar, Senegal. American Ethnologist 37(4), 638–661.

[73] Fullwiley, D. (2008) The Biologistical construction of race: 'Admixture' technology and the new genetic medicine. Social Studies of Science 38(5), 695–735.

[74] Arnedo, J., Svrakic, D.M., del Val, C. et al. (2014) Uncovering the hidden risk architecture of the schizophrenias: Confirmation in three independent genome-wide association studies. American Journal of Psychiatry in Advance, 172(2): 139–153.

[75] Townsend, J.M., (1980) Psychiatry versus societal reaction: A critical analysis. Journal of Health and Social Behavior 21(3), 268–278.

[76] Mirowsky, J., Ross, C.E. (1989) Psychiatric diagnosis as reified measurement. Journal of Health and Social Behavior 30(1), 11–25.

[77] Krumholz, H. (2011) Five lessons from niaspan's disappointing study. Available at: www.forbes.com/sites/harlankrumholz/2011/05/31/five-lessons-from-niaspans-disappointing-study/, accessed November 6, 2014.

[78] Harris, G. (2011) Study questions treatment used in heart disease. Available at: www.nytimes.com/2011/05/27/health/policy/27heart.html?_r=0, accessed November 6, 2014.

[79] National Heart, Lung, and Blood Institute (NHLBI) (2011) NIH stops clinical trial on combination cholesterol treatment: Lack of efficacy in reducing cardiovascular events prompts decision. Available at: www.nih.gov/news/health/may2011/nhlbi-26.htm, accessed November 6, 2014.

[80] Bodenreider, O. (2004) The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research 32, D267–D270.

[81] Mamykina, L., Miller, A.D., Medynskiy, Y., et al. (2011) Examining the impact of collaborative tagging on sensemaking in nutrition management. In Proceedings of the

SIGCHI Conference on Human Factors in Computing Systems, May 7–12, 2011, Vancouver, BC, Canada. ACM, New York, pp. 657–666.

[82] Mamykina, L., Greenblatt, D., Davidson, P.R., et al. (2008) MAHI: Investigating social scaffolding for reflective thinking, in diabetes management. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 6, 2008, Florence, Italy. ACM, New York, pp. 477–486.

[83] Mamykina, L., Miller, A.D., Mynatt, E.D. (2010) Constructing identities through storytelling in diabetes management. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 10–15, 2010, Atlanta, GA, USA. ACM, New York, pp. 1203–1212.

[84] Li, I., Dey, A.K., Forlizzi, J. (2011) Understanding my data, myself: Supporting self-reflection with ubicomp technologies. In Proceedings of the 13th International Conference on Ubiquitous Computing, September 17–21, 2011, Beijing, China. ACM, New York, pp. 405–414.

[85] Parker, A.G., Grinter, R.E., (2014) Collectivistic health promotion tools: Accounting for the relationship between culture, food and nutrition. International Journal of Human-Computer Studies 72(2), 185–206.

[86] el Kaliouby, R., Picard, R.W., Baron-Cohen, S. (2006) Affective computing and autism. In Bainbridge, W.S. & Roco, M.C. (eds.) Progress in Convergence. New York Academy of Sciences, New York, pp. 228–248.

[87] Sano, A., Hernandez, J., Deprey, J., et al. (2012) Multimodal annotation tool for challenging behaviors in people with autism spectrum disorders. In Proceedings of the International Conference on Ubiquitous Computing, September 5, 2012, Pittsburgh, Pennsylvania. ACM, New York, pp. 737–740.

[88] Escobedo, L., Nguyen, D.H., Boyd, L.A., et al. (2012) MOSOCO: A mobile assistive tool to support children with autism practicing social skills in real-life situations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, May 5–10, 2012, Austin, TX, USA. ACM, New York, pp. 2589–2598.

[89] Hourcade, J.P., Williams, S.R., Miller, E.A., et al. (2013) Evaluation of tablet apps to encourage social interaction in children with autism spectrum disorders. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 27–May 2, 2013, Paris, France. ACM, New York, pp. 3197–3206.

[90] Poh, M.-Z., Loddenkemper, T., Swenson, N.C., et al. (2010) Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 31–September 4, 2010, Buenos Aires, Argentina. IEEE, Piscataway, NJ, pp. 4415–4418.

[91] Huang, S.T.-Y., Sano, A., Kwan, C.M.Y. (2014) The moment: A mobile tool for people with depression or bipolar disorder. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, September 13–17, 2014, Seattle, WA, USA. ACM, New York, pp. 235–238.

[92] Rich, C., Sidner, C.L. (2010) Collaborative discourse, engagement and always-on relational agents. In Bohus, D., Horvitz, E., Kanda, T., et al. (eds.) AAAI Fall Symposium on Dialog with Robots, Technical Report FS-10-05. AAAI Press, Menlo Park, CA.

[93] Kalnikaitė, V., Whittaker, S. (2011) A saunter down memory lane: Digital reflection on personal mementos. International Journal of Human-Computer Studies 69(5), 298–310.

[94] Brynjolfsson, E., McAfee, A. (2011) Race Against the Machine. Digital Frontier Press, Lexington, MA.

[95] Frey, C.B., Osborne, M.A. (2013) The Future of Employment: How Susceptible Are Jobs to Computerisation? Oxford Martin Programme on the Impacts of Future Technology, Oxford.

[96] Czeskisy, A., Dermendjievay, I., Yapity, H., et al. (2010) Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In Proceedings of the Sixth Symposium on Usable Privacy and Security, July 14–16, 2010, Redmond, WA, USA. ACM, New York, #15.

[97] Friedman, B. (2004) Value sensitive design. In Bainbridge, W.S. (ed.) Encyclopedia of Human–Computer Interaction. New Barrington, MA, pp. 769–774.

[98] Bainbridge, W.S. (2015) The Meaning and Value of Spaceflight. Springer, London.

[99] Horgan, J. (1996) The End of Science. Addison-Wesley, Reading, MA.

[100] White, K.M., Preston, S.H. (1996) How many Americans are alive because of twentieth-century improvements in mortality? Population and Development Review 22(3), 415–429.

[101] Preston, S.H. (1996) American Longevity: Past, Present, and Future. Center for Policy Research, Syracuse University, Syracuse, NY.

[102] Bainbridge, W.S. (2007) Nanoconvergence. Prentice-Hall, Upper Saddle River, NJ, pp. 26–28.

[103] Bush, V. (1945) Science, the Endless Frontier. U.S. Government Printing Office, Washington, DC.

[104] Turner, F.J. (1920) The Frontier in American History. Holt, New York.

[105] Snell, J.E. (1967) Hypnosis in the treatment of the 'hexed' patient. American Journal of Psychiatry 124(3), 311–316.

[106] Darnton, R. (1970) Mesmerism and the End of the Enlightenment in France. Schocken, New York.

[107] Wolbring, G. (2003) Science and technology and the triple D (disease, disability, defect). In Roco, M.C. & Bainbridge, W.S. (eds.) Converging Technologies for Improving Human Performance. Kluwer Academic Publishers, Dordrecht, pp. 232–243.

[108] de Melo-Martín, I. (2010) Defending human enhancement technologies: Unveiling normativity. Journal of Medical Ethics 36(8), 483–487.

[109] More, M., Vita-More, N. (eds.) (2013) The Transhumanist Reader. Wiley-Blackwell, Chichester.

[110] Fukuyama, F. (2002) Our Posthuman Future: Consequences of the Biotechnology Revolution. Farrar, Strauss and Giroux, New York.

[111] Bostrom, N. (2005) In defence of posthuman dignity. Bioethics 19(3), 202–214.

[112] Stapledon, O. (1935) Odd John. Methuen, London.

[113] van Vogt, A.E. (1946) Slan. Arkham, Saux City, WI.

[114] Young, M. (1958) The Rise of the Meritocracy, 1870–2033. Thames and Hudson, London.

[115] Quine, W.V.O. (1959) Methods of Logic. Holt, New York.

[116] Popper, K. (1959) The Logic of Scientific Discovery. Basic Books, New York.

[117] Moore, G.E. (1903) Principia Ethica. Cambridge University Press, Cambridge, UK.

[118] Bainbridge, W.S. (1994) Values. In Asher, R.E. & Simpson, J.M.Y. (eds.) The Encyclopedia of Language and Linguistics. Pergamon, Oxford, pp. 4888–4892.

[119] Hitlin, S., Piliavin, J.A. (2004) Values: Reviving a dormant concept. Annual Review of Sociology 30, 359–393.

[120] Kant, I. (1981) Grounding for the Metaphysics of Morals. Hackett, Indianapolis, IN.

[121] Rawls, J. (1971) A Theory of Justice. Harvard University Press, Cambridge, MA.

[122] Homans, G.C. (1974) Social Behavior: Its Elementary Forms. Harcourt Brace Jovanovich, New York.

[123] Homans, G.C. (1964) Bringing men back in. American Sociological Review 29(6), 809–818.

[124] Parsons, T. (1964) Evolutionary universals in society. American Sociological Review 29(3), 339–357.

[125] Axelrod, R.M. (1984) The Evolution of Cooperation. Basic Books, New York.

[126] von Neumann, J., Morgenstern, O. (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

[127] Bainbridge, W.S. (2006) God from the Machine: Artificial Intelligence Models of Religious Cognition. AltaMira, Walnut Grove, CA.

[128] Wilson, E.O. (1971) The Insect Societies. Harvard University Press, Cambridge, MA.

[129] Wilson, E.O. (1975) Sociobiology. Harvard University Press, Cambridge, MA.

[130] Maynard Smith, J. (1982) Evolution and the Theory of Games. Cambridge University Press, New York.

[131] IEEE code of conduct. Available at: www.ieee.org/about/ieee_code_of_conduct.pdf, accessed November 19, 2014.

[132] Anderson, R.E., Engel, G., Gotterbarn, D., Hertlein, G.C., Hoffman, A., Jawer, B., Johnson, D.G., Lidtke, D.K., Little, J.C., Martin, D., Parker, D.B., Perrolle, J.A., Rosenberg, R.S. Code of Ethics and Professional Conduct. Available at: www.acm.org/about/code-of-ethics, accessed November 19, 2014.

[133] Gornall, J. (2007) Hyperactivity in children: The Gillberg affair. British Medical Journal 335(7616), 370–373.

[134] Markham, A., Buchanan, E. (2012) Ethical decision-making and internet research. Association of Internet Researchers, p. 3. Available at: www.aoir.org/reports/ethics2.pdf, accessed November 28, 2014.

[135] Wikipedia, Mug shot publishing industry. Available at: en.wikipedia.org/wiki/Mugshots_removal, accessed October 29, 2014.

[136] Ratkiewicz, J., Conover, M., Meiss, M., et al. (2011) Truthy: Mapping the spread of astroturf in microblog streams. In Proceedings of the 20th International Conference Companion on World Wide Web, March 28–April 1, 2011, Hyderabad, India. ACM, New York, pp. 249–252.

[137] Ratkiewicz, J., Conover, M., Meiss, M., et al. (2011) Detecting and tracking political abuse in social media. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, July 17–21, 2011, Barcelona, Spain. AAAI Press, Menlo Park, CA, pp. 297–304.

[138] Sieghart, P., Dawson, J. (1987) Computer-aided medical ethics. Journal of Medical Ethics 13(4), 185–188.

[139] Mizani, M.A., Baykal, N. (2007) A software platform to analyse the ethical issues of electronic patient privacy policy: The S3P example. Journal of Medical Ethics 33(12), 695–698.

[140] Shankar, G., Simmons, A. (2009) Understanding ethics guidelines using an internet-based expert system. Journal of Medical Ethics 35(1), 65–68.

[141] Wallach, W. (2012) Mechanizing ethics. In Bainbridge, W.S. (ed.) Leadership in Science and Technology. Sage, Thousand Oaks, CA, pp. 149–156.

18

THE ETHICS OF UBIQUITOUS COMPUTING IN HEALTH CARE

Clark A. Miller, Heather M. Ross, Gaymon Bennett, and J. Benjamin Hurlbut

School of Politics and Global Studies, Arizona State University, Tempe, AZ, USA

18.1 INTRODUCTION

Ubiquitous computing in health care raises significant ethical challenges that must be confronted as part of any comprehensive effort to understand the technological and institutional reconfigurations of this emerging domain. Some problems of ethics in ubiquitous computing in health care follow in line with ethical problems already well addressed in medical, health, and bioethics. The principal product of ubiquitous computing technologies, after all, is greater information about the health of patients. Health information has been the subject of extensive ethical analysis, especially with regard to concerns about information privacy, the autonomy of patients, and the use of health information in discriminatory practices [1]. Considerable ethical work has been done on issues of privacy, autonomy, and discrimination in genetic medicine in particular, which represents an early threshold in the deep integration of computing and computational analysis in health care [2]. We will argue, however, that while such concerns remain significant, the ethical challenges posed by ubiquitous computing in health care go beyond these issues. Indeed, these challenges complicate and, in some respects, transform existing notions of risk and privacy with regard to information.

We argue that the widespread adoption of ubiquitous computing in the healthcare sector raises distinctive ethical challenges because it portends the transformation of

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

not just health technologies but also the broader healthcare systems within which those technologies are embedded. The effects of ubiquitous computing will ripple outward bringing subtle but significant shifts in the day-to-day practices of medicine, its financial models, institutional organization, and relationships among key actors (e.g., patients, consumers, healthcare providers, clinics, insurers, regulators, scientists, and device manufacturers), as well as broader concerns of medical science, expertise, and authority. As these ripples reconfigure and transform the socio-technical arrangements of contemporary health knowledge and institutions, they will raise important questions about the ethical design of health systems and the ethical distribution of outcomes that flow from them. It bears noting that the potential impacts of ubiquitous technology tend to be imagined as running in one direction, as though the capacities of new devices and techniques are sufficient to initiate wholesale reconfigurations of medical practice. We think that such a vision overstates things. Rather, existing medical infrastructures and normative regimes, which in many ways are deeply entrenched and stabilized, will equally shape the forms and functions that ubiquitous technologies take. However, we will not explore this dimension of the current situation in this chapter.

In this chapter, we use the term "ubiquitous computing" to refer to an array of diverse technologies deployed throughout multiple stages of healthcare delivery, from research and development, to patient care and therapeutics in health institutions, to the monitoring and surveillance of health and health inputs in everyday living. These include, for example, the following: (i) Web-enabled, smart medical devices that may be either surgically installed inside the body or placed on the body; (ii) data communication between those devices and the patient, other medical devices and instruments, for example, in surgeries or hospital rooms or in doctor's offices, the patient's electronic health records, healthcare providers, insurers, artificial intelligence agents, and/or scientific research laboratories; (iii) capabilities for remote intervention in device operation; and (iv) information systems and infrastructures for collecting, storing, aggregating, synthesizing, analyzing, and communicating the resulting "big data" to inform patient care, biomedical research, public health surveillance, and innovation in bioinformatics technologies.

The following three related but distinct visions are driving the rapid growth of ubiquitous computing in health care:

1. *The knowledge-empowered individual*: In this vision, new devices that sense and collect information about the body, habits, and activities promise the individual access to dimensions of health that were previously out of reach, either because technologies for data collection did not exist or were controlled by healthcare providers. These technologies are taken as offering new forms of autonomy and individual power over health, even as they create new dependencies on the technologies that mediate these experiences and new imperatives to manage one's own health.
2. *Routine health surveillance*: In this vision, similar health technologies (often with subtle but important differences in socio-technical design, which we define as the combination of technical design elements alongside design of

ways that people interact with the technology, for example, with regard to its financing, the rules governing its operation, or access to the information it provides) provide healthcare institutions with new abilities to collect, aggregate, store, process, and analyze data about users. These data can be communicated and used within the health system in order to coordinate diagnosis, prediction, and response to therapy.

3. *Digital medicine*: The third vision focuses less on the interests and information needs of patients and healthcare providers and more on the power of information technologies to transform medical research by creating new information infrastructures that will, in cardiologist Eric Topol's words, "digitize the human." In the process, information technologies may profoundly reorder medical understanding of health and disease, thereby transforming knowledge; clinical care; and the economic, institutional, and regulatory structures of biomedicine [3].

We describe these visions in greater detail in the next section, as they shape the design of new health systems animated by ubiquitous computing—sometimes aligning, sometimes competing—and thus are important facets in shaping its ethical dimensions. In practice, the lines between these visions are blurry. That blurriness accounts for some of the ethical tensions and contradictions that are evident in this emerging domain.

In each of these visions, ubiquitous computing is imagined in terms of extensive infrastructures; it is not a collection of one-off devices, but an interlinked web. The functionality of any device is dependent on—and integrated into—larger infrastructures of data collection, aggregation, and analysis. The alignment of these three visions has justified huge investments in these infrastructures. Many of the world's largest consumer, information, and medical device companies are investing enormous financial and human capital in the development of new health information devices and technologies, information processing and communications infrastructures, and "big data" health analytics. Major health institutions of all sorts are positioning themselves as centers for the collection and aggregation of patient data. Start-ups abound in all facets of this arena. The US government has financed massive new information enterprises for biomedical research and the standardization and adoption of electronic health records throughout the medical industry.

Ethically, the rapid development of these new infrastructures is highly significant for three reasons. First, the entrance of ubiquitous computing into health care signals the joining of two existing socio-technical systems. Even as personal digital devices are being connected to the medical body, the medical body is being connected into the emerging Internet of Things, an interconnected web of devices embedded in common objects in environments of daily life that communicate without requiring human–human or human–computer interaction.

Second, the development of these infrastructures affects the lives of patients, the practice and organization of medicine, and the pursuit of medical research—long before there is strong evidence of the promised transformative improvements in health. Indeed, the very possibility of transformational outcomes depends on these

infrastructures being securely in place and operating for a sufficiently long period of time for scientists, doctors, insurers, regulators, policy-makers, businesses, and individuals to learn how to make use of the resulting information to improve health. This dilemma poses serious ethical questions about resource allocation for public health (e.g., between future promises and improved medical care today), as well as about the meaning and implications of these infrastructures for the lives of those who inhabit them, especially patients and healthcare professionals who may find themselves confronting new forms of unexpected vulnerabilities.

Third, these infrastructures are intrinsically transformative, whether they improve health or not. Their transformative characteristics, however, remain underdetermined; they are flexible and not (yet) fixed. The three visions we discuss diverge in the design of the socio-technical arrangements and infrastructures that they envision as necessary for their accomplishment and thereby set into motion. This divergence is partly a consequence of the different ideas of harm and benefit at work in each vision, and the divergence itself creates ethical dilemmas. For example, consumers who wish to reduce their dependence on health institutions by using smart technologies for self-testing, self-diagnosis, self-interpreting, and self-treatment—an ambition of many in the "quantified self-movement"[1]—may find that they have to relinquish ownership and control over their data so that it can be aggregated and analyzed in health information architectures that enable it to provide meaningful insights. It also creates conflicts over proper system design, often with deep ethical dimensions. Patients who think that greater information may give them more control over their own health decisions may find, for example, that their access to information from medical devices is highly limited, either by device companies eager to maintain information monopolies, by doctors reluctant to hand information over to patients without medical expertise, or by regulations that do not acknowledge a right to data access. Ethical analysis, therefore, confronts a dual challenge: to assess the emergent characteristics of newly forming information infrastructures contributing to the development of ubiquitous computing in health care and, just as important, to help create appropriate design standards for those infrastructures to be judged ethical.

Whatever directions ubiquitous computing takes, one can imagine that ubiquitous computing will develop in multiple directions, both foreseen and unforeseen. The potential outcomes (health, economic, and otherwise) raise critical ethical questions that must be posed and reposed as ubiquitous computing is integrated (or not) into healthcare institutions and practices. Vigilant and critical attention to ethical questions is crucial to being able to direct and govern the development and adoption of new technologies in a fashion that attends to potential harms as well as potential goods. In what follows, we lay some of the groundwork for this exercise. First, we describe in more detail the three prominent visions driving the development and adoption of ubiquitous computing in health care and their connections to diverse socio-technical and ethical imaginations of progress in biomedicine. Second, we provide a short case study of one of the most significant established domains of the application of smart

[1] It broadly comprises individuals who use a variety of devices to measure and analyze multiple aspects of their daily living, with the goal of learning about and improving their health and physical performance [4].

medical devices in the field of cardiac care and its emergent forms of ethical dilemmas and challenges. Third, we develop a more layered analysis of the ethics of ubiquitous computing in health care across a range of diverse applications and developments. Finally, we conclude with a few modest suggestions for strategies that seek to energize deliberations among experts and publics concerning the ethical significance of new directions in the design of health information infrastructures.

18.2 UBIQUITOUS COMPUTING AND THE TRANSFORMATION OF HEALTH CARE: THREE VISIONS

Apple's Health app is integrated into the iOS operating system for the iPhone, networked to a rapidly growing suite of third-party health, nutrition, and fitness apps, and linked to a heart-rate sensor built into the Apple Watch. Advertising for the Health app epitomizes one prominent vision of the future of ubiquitous computing in health care. As noted in the introductory section, we term this vision *the knowledge-empowered individual*. In this vision, ubiquitous computing empowers each individual to take much greater control of his/her health through a knowledge system that generates new and better information, synthesizes it, and makes it possible to use the information to improve health decisions, from daily choices about eating and exercise to monitoring of a variety of short- and long-term health conditions.

The vision of the knowledge-empowered individual focuses on the proliferation of devices designed to sense environments, behaviors, and physiological and psychological states to provide the user with better information about his/her own health. It is underwritten by the notion of an active, engaged individual with personal agency empowered by technologies for collecting and using otherwise inaccessible information about his/her own body to monitor and improve his/her own health. For example, smartphones and tablets already sense and record health-related information about the users who carry them. Sensing, computing, and communicating devices are likewise present in a growing number of everyday spaces from households to schools to corporate campuses. The network of devices that facilitate ubiquitous computing can be found in automobiles, appliances, cameras, headphones, wristbands, and shoes, sometimes referred to as the Internet of Things. Ubiquitous computing—facilitated by the diffusion of these Internet-embedded computing devices—makes it possible for medically pertinent data to be extracted from virtually all dimensions of our everyday lives and activities and presented to us in increasingly convenient and persuasive ways. Apple's HealthKit offers a good example:

> The new Health app gives you an easy-to-read dashboard of your health and fitness data. And we've created a new tool for developers called HealthKit, which allows all the incredible health and fitness apps to work together, and work harder, for you. It just might be the beginning of a health revolution … "How are you?" now has a really accurate answer. Heart rate, calories burned, blood sugar, cholesterol—your health and fitness apps are great at collecting all that data. The new Health app puts that data in one place, accessible with a tap, giving you a clear and current overview of your health … You are in charge of your data.

> The Health app lets you keep all your health and fitness information in one place on your device and under your control. The information you generate about yourself is yours to use and share. You decide what information is placed in Health and which apps can access your data through the Health app. (apple.com November 13, 2014)

Apple's marketing vision encapsulates and caters to the sensibilities of the *quantified self-movement*. Yet, Apple's system is, in reality, far more powerful, and the power of the individual to make use of health data implied by Apple's marketing rhetoric is in tension with the ability of app developers to aggregate and sell products based on an individual's health data. For example, alongside the Health app, which is designed to aggregate and store health data, is HealthKit, a set of developer tools that allow other apps to access data stored in Health app and use the data for other purposes. Apple stresses that strict privacy controls will allow the individual to retain control over what information is shared and with whom. However, the key to HealthKit is the ability to create a virtual health data ecosystem[2] in which many companies are able to make money by selling devices, software, information, and expert interpretations of health data to Apple customers. The ecosystem is predicated on the assumption that the individual will not exercise his/her right of control. Apple recognizes the conflict between these two positions and has developed a set of guidelines[3] designed to encourage app developers to approach this task in ways that foster the trust of users. However, such guidelines do not remove the tension, because the user is ultimately less in control of their data than the vision of self-empowerment implies. Or, rather, the vision of self-empowerment requires a user who is empowered by surrendering his/her data in exchange for new ways of analyzing it.

This more ambitious yet ambiguous model of health information sharing leads to the second major vision of ubiquitous computing in health care: *routine health surveillance*. In this vision, the healthcare sector uses various smart medical devices to continuously monitor patient health and health behavior. If Apple's new Health app and sensors on the iPhone and Apple Watch epitomize the knowledge-empowered individual model, routine health surveillance is perhaps best represented by the growing capabilities of "smart" medical devices, such as pacemakers, which now systematically collect data and wirelessly transmit that data to doctors and clinics with no intervention from the patient.

The routine health surveillance model renders the individual patient's body visible and readable to the medical practitioner to a greater extent and in new ways. Equipped with these tools, doctors will no longer need to rely on periodic diagnostic tests or patient reports but rather will have access to continuous data records that will greatly facilitate diagnosis and monitoring of their patients' health and health determinants.

[2] By virtual health data ecosystem, we refer to the emerging ensemble of data technologies, practices, and content that allows for the collection, integration, aggregation, circulation, interpretation, and leveraging of individual health data into the kind of large-scale health databases envisioned by the proponents of precision and personalized medicine.

[3] Apple's Human Interface Guidelines can be found at: https://developer.apple.com/library/ios/documentation/UserExperience/Conceptual/MobileHIG/index.html#//apple_ref/doc/uid/TP40006556

Ubiquitous computing technologies are used in this sense to extend existing forms and logics of clinical medicine, continuing existing forms of monitoring but significantly increasing their frequency, expanding the number and kinds of sensors, and extending sensor use into the patient's home life. Thus, existing clinical relationships are preserved. Medical professionals remain largely responsible for collecting, interpreting, and acting on information. Deployed in this way, ubiquitous computing will intensify existing regimes of expertise and control that characterize medical professionals (e.g., lab techs, actuaries, physicians, nurses, and biomedical researchers) and institutions (e.g., diagnostic laboratories, insurance companies, clinics, and research centers). The fine-grained and near real-time access to an individual patient's data, and the ability to interface that data with other information about that patient along with analytically aggregated data from hosts of other patients, will allow professionals and institutions to exercise greater control over patients and patient health. Such control generates new capabilities to regulate access to care and to conduct routine surveillance of patient health information as well as patient behaviors that contribute to health. At the same time, the potential for greater control also implies heightened responsibility. For instance, constant access to patient data implies a duty to constantly monitor, interpret, and respond when data are medically actionable. In short, the array of networked sensors, artificial intelligence agents, data sets, and remote health intervention capabilities may intensify the ability—and responsibility—of actors in the existing health regime not just to know about patient health but also to shape and control the physiological, behavioral, and social factors that contribute to it.

The routine health surveillance vision aims to promote patient health, like the vision of the *knowledge-empowered individual*. But it does so via a very different logic. Rather than empowering individuals to make good health decisions, *the routine health surveillance* vision rests primarily on the ability of the healthcare industry to properly monitor patient bodies, detect diseases, and other biological failures, and intervene to provide corrective action—whether to remind the patient to take his/her medicine or to undergo surgery to correct a faulty heart muscle. The surveillance model exaggerates a number of existing features of contemporary healthcare systems. Medical testing, which is largely under the control of the patient's healthcare provider and the laboratory conducting the test, has become a major component of the healthcare industry and increasingly provides not only one-off diagnostic tests, used to detect disease onset, but also a widely growing array of periodic health surveillance. Today, monitoring services are most often provided on a periodic basis associated with regular visits to a provider's office. Ubiquitous health computing offers the potential to make monitoring frequent, and perhaps continuous, and divorce it from the visit to the clinic.

Consider glucose monitoring for diabetes, for example. Glucose monitoring is a well-established market that, until recently, relied almost exclusively on "finger sticks" in which a patient took a blood sample with a small needle and then ran it through an assay to measure blood glucose levels. Visit the website of a company like Medtronic, and you will find a much different world evolving. A variety of products are now offered for "continuous glucose monitoring" that routinely monitor blood

glucose levels by inserting an electrode under the skin. The device then transmits data wirelessly to a monitoring device that collects and visualizes the data for the patient, or in the case of children, the patient's parents. While such devices fit comfortably into the knowledge-empowered individual paradigm, enabling diabetics to upgrade their ability to monitor and control their own blood glucose levels, smart devices increasingly also supplement patient self-monitoring with both artificial expert diagnoses and clinician surveillance. For example, Medtronic sells what it calls the Guardian Continuous Glucose Monitoring device, a tool that anticipates trends in blood glucose and provides predictive alerts when problematic blood glucose levels are approaching, based on real-time monitoring of glucose patterns. Much as the Nest thermostat learns a household's preferred patterns of home heating and cooling, so the Guardian learns a patient's regular glucose levels and can use them to better control diabetes. Finally, Medtronic sells CareLink software that makes the resulting data from its monitoring devices available to healthcare providers, so that they can provide routine oversight between visits to their office and no longer need to rely on vague self-reporting from patients about how well they're doing in controlling their blood glucose levels. Instead, providers can obtain the data directly from the device.

Another example of the growing trend toward routine health surveillance via smart medical devices is the AliveCor heart monitor that provides more information about a patient's heart rhythm health than can typically be obtained via periodic office visits. The AliveCor monitoring system also engages additional experts unrelated to the patient's established healthcare team, to provide interim interpretation of heart rhythm data obtained outside of a usual clinic visit.

> I now have a more complete view of what's happening with my patients between three or six month appointments, specifically their AF episodes and medications. My patients can also send me this information quickly, even during an AF episode, helping me give them immediate advice about what to do next. [statement by a doctor who uses AliveCor]. (alivecor.com November 13, 2014)

Smart medical devices such as the AliveCor monitor are already integrated into medical practice. However, as we will see in the next section, which examines the history of their clinical adoption, the ultimate ramifications of their use remain uncertain.

The third vision of ubiquitous computing in healthcare, *digital medicine*, de-emphasizes the interests and information needs of patients and providers in favor of an enhanced capacity for biomedical research. By eliciting large volumes of data about people's bodies and behaviors, proponents of *digital medicine* imagine that new health devices will bring about a transformation in medical knowledge. Throughout biomedical research and the healthcare industry, vast new data collection initiatives are under development or in process, to take advantage of the knowledge provided by new health information devices and other information technologies, such as genetic sequencing, biobanks, and high-throughput biomarker detection. Such technologies promise new insights for biomedical scientists and health

institutions that can take advantage of the data—whether to improve patient care, advance market positioning, or achieve some other significant objective. In this *digital medicine* vision, ubiquitous computing technologies would serve as a conduit for mining health data from individuals: data that may serve no immediate clinical purpose but would serve as the basis for applying "big data" analytic techniques to biomedical innovation [5, 6]. A report from the Institute of Medicine outlines how the convergence of biomedical science and information technology will generate extensive bodies of data about human health from which a "new taxonomy of disease" will emerge [5]. The report predicts that diverse technologies for eliciting data from and about human bodies will generate massive repositories of information that will reveal dimensions of human physiology, behavior, and environment that were previously unknowable, giving radically more precision to understanding of the causes of and effective interventions in human health and disease.

By layering genomic, proteomic, physiological, behavioral, environmental, and clinical data, it is expected that scientists will be able to fashion a so-called precision medicine that will become the basis for clinical decision-making: what the report describes as "Google maps" for health. By mapping disease in multiple dimensions, the *digital medicine* vision promises an easy and straightforward path to navigate to better care and superior health outcomes. In this vision, ubiquitous medical computing is a fundamental infrastructural element. Global mapping of health and disease will require a "street view" surveillance of individual human bodies. Such a task will depend on mobile health technologies on or in those bodies, as well as a host of other technologies, from genetic sequencing to biobanks. In order to make this map, medical populations must undertake—and become the subject of—a massive cartographical project. Even bracketing the question of whether this mapping exercise will lead to better health outcomes, constructing the technological regimes that can generate this information raises difficult issues about the relationship (and tensions) between obligations to the individual and the imperative of advancing medical knowledge, between the promotion of individual and public health, and between improving health research and improving health.

Precisely how the future of ubiquitous computing in health care will play out remains highly uncertain. One of these three visions may dominate, a mix of all three may persist, or entirely new configurations may emerge. For example, it is easy to imagine a world in which insurance companies become a primary consumer of health information from devices. Today, insurers rely heavily on clinician judgments regarding patient health, but they may in the future seek to take advantage of more direct health measurements. Such measurements would be extremely valuable to health insurers in at least three ways: to control and regulate access to health services; to assess whether patients are behaving healthily, and therefore are eligible for health coverage; and to monitor whether patients are following through on health treatments, such as taking their medicine, again, as an input to coverage eligibility and payment for services.

We discuss these visions for two reasons. First, they reveal that health innovation is not the consequence of technological change alone, but always entails corollary institutional and economic changes in the organization and practices of the health

system. In other words, what matters is what kind of future health, health systems, health economies, and norms of care are brought into being in conjunction with new technologies—not merely the future technologies per se. Importantly, this implies that innovation at the systems level is not determined by technological developments alone. Systemic innovation can occur whether or not a given technological vision is fully realized, since innovation agendas themselves involve normative reorientations. They embody visions of the way medicine ought to be and assert altered forms and allocations of responsibility and resources, even as the technological infrastructure that underwrites these alterations remains aspirational and emergent. Second, these possible configurations raise profound ethical questions that deserve careful treatment and consideration in the design of future health arrangements surrounding ubiquitous computing. Who should and will control health data and its interpretation? How should and will responsibilities of diverse participants in the health system shift? How should and will the costs of ubiquitous computing shape healthcare system priorities, with what impacts on health outcomes and exclusions? How should and will the risks and benefits of ubiquitous computing be distributed across different groups? How, in short, can the technical and medical capacities of ubiquitous computing be institutionalized and increased without also intensifying existing disparities in empowerment, value, and control.

18.3 CASE STUDY: CARDIAC IMPLANTED ELECTRICAL DEVICES

In contemplating the ethical challenges raised by ubiquitous computing technologies in health care, it is useful to examine a case study that explores how such technologies are beginning to play out in clinical practice, to illuminate in real-world settings some of the ethical issues that arise with a transition from episodic to ubiquitous data communication. Cardiac implanted electrical devices have been a part of the medical landscape since the first pacemaker was implanted in Sweden in 1958 [7]. The first pacemaker, which lasted for 3 h and was the size of a hockey puck, has given way to much smaller implantable pacemakers with an expected lifespan of more than a decade and cardioverter-defibrillators that function nearly as long. Alongside these long-lasting technologies, significant improvements in battery technology and microcircuitry make possible sophisticated programming functions. The device can mimic or replace normal heart rhythm and provide detailed monitoring, capture, and reporting of data describing the functioning of the device technology and the patient's health status.

Typical cardiac implanted electrical devices record and report data related to heart rate, heart rate variability, abnormal rapid heart rhythms (including the chamber of origin and electrographic tracings), physical activity and movement, and even intrathoracic fluid status in some advanced devices. These data can be used by heart rhythm professionals to compile a detailed portrait of a patient's heart rhythm and heart failure status between episodic visits to the clinic. For patients who have a cardiac implanted electrical device, the narrative response from asking, "How have you been since your last appointment 6 months ago?" can be supplemented with data

to quantify the patient's qualitative report. For example, "I had a week in June where I had no energy, then it just got better," might be supported by a recorded episode of atrial fibrillation that lasted for a few days during that time period. Or, a patient who reports trouble losing weight despite rigorous daily exercise might show no activity on their pacemaker's accelerometer, leading the clinician to question the patient more carefully about their exercise regimen.

Since 2001, heart rhythm clinicians have been able to remotely access cardiac device data using specialized hardware and software to "interrogate" or extract data from the device's memory. Remote transmissions result in a transfer of technical and physiological data measured by the device to a secure website for access, analysis, and interpretation by the patient's care team. From 2001 to 2005, this functionality required discrete action by the patient to initiate data transmission. However, beginning in 2006, device manufacturers began to incorporate radio-frequency signaling technology into implantable devices [8].[4] This allowed remote communication independent of human initiation of the transmission. Heart rhythm clinicians can use an Internet-based system to schedule a remote transmission without directly communicating with the patient. Also, specialized algorithms in the cardiac devices can initiate a transmission in the event of a detected arrhythmia, even without the patient or clinician being aware of the arrhythmia or the "unscheduled" transmission [9].

These remote data transmissions are unidirectional, with data uploaded from the patient's device accessible only by the patient's designated clinicians. Patients do not have direct access to their recorded data from implanted cardiac devices unless the clinician provides it to them specifically. Device companies reinforce this unidirectionality by claiming that physicians and hospitals are their customers, not patients. Those same companies, moreover, report that federal guidelines prevent them from sharing data directly with patients without obtaining approval from the US Food and Drug Administration (FDA) for that function.

Present US standards for the chronic care of implanted cardiac devices require regular device interrogations with hardware testing every 3–6 months for most patients. In person, this requires the patient to schedule an appointment in the clinic, check in for the appointment with clerical staff, wait until the clinician is ready, undergo device interrogation and testing for approximately 5–20 min, and potentially wait for another clinician to address any problems that are uncovered during the interrogation and testing. This entire process—if uncomplicated—requires approximately 30 min of staff time (including pacemaker clinician and office staff for check-in/check-out processes) and occupies an exam room for 20 min or more. Cardiology clinics are typically designed and staffed for this work during the routine business day.

Remote monitoring of implanted devices has allowed patients and clinicians to replace several in-person office visits each year with remote transmissions, ostensibly saving both patient and clinician time and money by reducing the need for costly

[4] Device manufacturers with radiofrequency signaling technology for remote monitoring include Medtronic (www.medtronic.com), Biotronik (www.biotronik.de), Boston Scientific (www.bostonscientific.com), St. Jude (www.sjm.com), and Sorin (www.sorin.com).

travel and staff time and equipment involved with live visits. With the integration of remote monitoring, many cardiology clinics have shifted some worker responsibilities to accessing, interpreting, and processing the data from these so-called remote visits, rather than focusing full-time attention on patients in the physical clinic. In a typical arrangement in the present age of electronic medical records, a technician enters the remote device data into the medical record, and generates a report for the clinician to view, modify, and approve electronically. The clinician is prompted, typically in a software inbox or dashboard, that a report is awaiting review.

In one busy suburban private practice, four-physician, electrophysiology group, approximately one full-time equivalent (FTE) specialist cardiac device technician is dedicated to downloading, analyzing, and processing remote device interrogation data into the electronic medical record for heart rhythm provider evaluation, documentation, and communication with other members of the patient's care team (e.g., referring cardiologist, internist, and other relevant specialist clinicians). This process takes an additional 0.2 FTE of a specialist cardiovascular clinician (e.g., physician, nurse practitioner, or physician assistant).

In this particular private practice, patients with pacemakers and defibrillators that function normally and do not require frequent in-person adjustment experience remote interrogations every 3 months and have an in-person office visit just once a year, compared to quarterly in-person visits without remote monitoring. These routine remote interrogations are scheduled on Monday mornings, so that the technician can anticipate and attend to the transmissions shortly after they arrive on the internet dashboard. Patients with manual remote systems that require patient activation of the download are instructed to perform the download anytime on Sunday or on Monday morning. Devices with wireless remote systems are scheduled to automatically download on Monday mornings. Therefore, on Mondays beginning at noon, the technician is able to perform a rapid scan of the transmissions to assess for any troublesome findings, then can proceed through each monitored report individually to catalog and report the findings. Similar monitoring workflows are in place in large urban academic medical centers and small rural clinics alike.

Patients who experience troubling symptoms are instructed to notify the practice prior to performing an unscheduled or nonroutine transmission. This phone communication alerts the technician to incoming data. However, patients often send an unscheduled transmission without notifying the device clinic, leaving the technician with no direct alert to anticipate incoming data. Unscheduled transmissions are particularly troublesome when the patient may be experiencing significant distress as a result of abnormal heart function or malfunction of the implanted device, but data may not be recognized in a timely manner. This raises serious questions about both the imagined timeliness of real-time data and the expectations that a digitally powered network of connections will improve reaction time, accountability, and patient outcomes.

In addition to standard Monday transmissions, technicians in this practice scan the remote monitoring dashboards several times during each workday to check for additional unscheduled transmissions. These are evaluated and triaged to healthcare providers for additional evaluation as needed. In some cases, these transmissions

yield a brief letter or phone call to the patient. However, sometimes the transmissions signal the need for in-person evaluation either in the clinic or by emergency services. Such urgent cases often disrupt the already busy flow of the clinic schedule. Sometimes, routine patient appointments must be canceled or rescheduled in order to accommodate the urgent situation. This is costly for clinical staff and patients alike, including indirect costs related to transportation, missed work, and associated care requirements for frail patients.

Wireless remote monitoring raises similar, and in some ways more troubling, concerns to unscheduled device transmissions. With wireless remote monitoring, data transmissions may occur without even the patient being aware. This means that the device might transmit information about a hardware failure or a concerning arrhythmia without any direct intention from patient or clinician. Contemporary wireless remote monitoring systems afford clinician notifications by website, voice messaging, e-mail, text messaging, paging, or a live phone call for certain "red alert" conditions. These notification preferences and alert-level designations are programmable and modifiable by the clinician on a clinic-wide or per-patient basis. Designating notification preferences includes making decisions about workflow as well as judgments about timely accountability.

Wireless alert notifications require practices to internally designate protocols for receiving alert notifications. For example, if a practice designates a particular technician to primarily monitor the clinic's Internet dashboards for each device company's remote monitoring data, how often will these sites be scanned to search for new wireless alerts? In the present landscape with five separate manufacturers of implanted cardiac devices, each with separate, proprietary remote monitoring systems, how should multiple sites be simultaneously monitored? Will the remote monitoring system websites be monitored outside of business hours and on weekends? Who will be responsible for after-hours monitoring (and what will be the implications for hourly or salaried employees to take on additional monitoring responsibilities)? What happens in the case of system failures, scheduled downtimes, or when the primary designee is sick or on vacation? These same questions are also relevant for practices that elect to have more direct notification of alerts via e-mail, text messaging, or paging. The increased monitoring requirements may dictate the need for practices to hire and train scarce and costly specialist staff without any anticipated reimbursement to compensate the additional expense.

Ubiquitous monitoring expectations also raise questions about the expected standards of accountability and responsibility for addressing and acting on data. What is the expected response time for acting on an alert after notification? In the event that an alert is set for delivery via SMS message, what is the clinician's responsibility if the target cell phone is out of range or there is another technology failure? At present, remote transmission of patient data is legally subject to privacy protections in the United States according to the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Other aspects of remote data monitoring, however, lack such clear legal standards. Interstate practice regulations are unclear, and there is no US or European legal standard for clinician monitoring expectations in the event of an adverse outcome.

Clinicians must also make decisions about which conditions merit alert-level notification. These decisions have implications for generating errors of sensitivity or specificity, based on prioritizing data communication. For example, should an alert be delivered if a patient has atrial fibrillation lasting for more than 4 h, or should clinicians be alerted only for episodes lasting more than 24 h? Either condition raises the risk of a stroke, particularly if a patient is not taking blood thinner medications. However, the relative risk of a stroke must be prioritized against the burden of data management by healthcare workers, and the possibility of causing alarm fatigue with many alert notifications. The process of designing notification criteria also raises issues related to autonomy and individualized care. Should each practice designate a "standard set" of conditions meriting after-hours notification? How and when should these standards be modified for a particular patient? Should individual patients be involved in the decision-making process for setting alert standards? These are challenging questions of standard setting in their own right, but we raise them also to demonstrate how ubiquitous computing technologies that ostensibly merely extend existing forms of monitoring actually represent quite significant perturbations to existing standards and practices with significant ethical implications. The fact that such perturbations can result from a technology that generates information that does not differ in kind but merely in volume, frequency, or routine availability suggests that these challenges may well become more complex where technologies elicit new forms of information that are not yet routinely employed in clinical care.

At present, a typical clinic operates on a 24/7 basis only for directed emergencies triggered by individual human contacts to the practice (e.g., a phone call to the after-hours answering service or a visit to the emergency room). If clinicians become responsible for constant monitoring of continuous data streams at significant cost, then who will be responsible for incurring that cost? Should modified alert algorithms be available to patients at an extra charge from the typical monitoring fees that clinics receive from insurance companies? On a larger scale, continuous monitoring would require significant inputs of specialized healthcare personnel time and expertise. In the current climate characterized by a supply–demand mismatch in terms of expert physician and non-physician providers compared to population health needs, ubiquitous medical computing and data management raises concerns about further reducing access to limited healthcare expertise for other patients and other aspects of patient health [10, 11].

Similarly, the expectations for reacting to continuous medical data raise ethical questions about imposing new burdens on emergency resources. If the transmission capability is instantaneous, should it be treated as an emergency notification system? If a clinician receives an alert in the midnight for a patient, should she contact the patient immediately or wait until morning? How many times, or for how long, should the clinician attempt to contact the patient? If a clinician attempts to contact the patient by phone but is unsuccessful, should scarce public emergency resources (e.g., fire rescue) be alerted and sent to the patient's home? In the event of a bad outcome for the patient (i.e., discomfort, injury, or even death), what is the clinician's legal liability based on remotely transmitted cardiac device data? In effect, the sheer fact that data exist raises challenging questions about "reasonable expectations" on

the part of patients, regulators, the media, and publics for the responsibility of the healthcare system. What expectations are reasonable in turn depends on how devices are seen as integrating into healthcare. Whereas the above scenarios presume a *routine heath surveillance* paradigm, the allocations of responsibility are fundamentally different if data are treated as a consumer good acquired by a *knowledge-empowered individual*. In this sense, information is never neutral, but is always embedded in prior allocations of responsibility. Position papers from professional bodies have raised the need to establish meaningful standards and oversight with regard to remote monitoring responsibility (e.g., Refs. [12, 13]). As of this writing, however, no such meaningful standards exist.

Such stakes are particularly evident in complex ethical issues about data ownership and access to medical data. Questions of who should own and have access to data also implicate contested notions of who is—and should be—empowered to take responsibility for individuals' health. In 2013, one self-declared "e-patient" gained notoriety for publicly demanding access to the data generated by his cardiac device. Posting to the Twitter feed of Medtronic CEO Omar Ishrak, Hugo Campos asserted that Medtronic was exerting illegitimate control over information that rightly belonged to him because it was from and about his own body. In response to a post by Ishrak about using a health app to monitor his exercise, Campos responded "thumbs up for using #mHelath apps to track progress. Thumbs down for not giving same access to patients with your devices. #FAIL." Campos' post elicited a string of comments rebuking Medtronic for not giving patients access to the data derived from medical devices. Medtronic responded by making a special exemption and writing a script specifically for Campos that would generate a spreadsheet with the raw data every time Campos initiated a transmission. Medtronic installed this script in Stanford University Hospital's system where Campos was a patient. Stanford then billed uninsured Campos nearly $800 dollars for providing him with the two-page-long spreadsheet of raw data [14]. Campos's intervention has been celebrated as an example of a would-be empowered patient pushing back against an established system of routine medical surveillance that treats data as proprietary and thus precludes patients' access to their own data. It reveals a fracture between competing visions of what—and whom—these data are for, and what norms should govern ownership and access. And it highlights the quite fundamental ethical perturbations to medical relationships posed by ubiquitous medical computing.

18.4 ETHICAL REFLECTIONS

From the perspective of medical ethics, and bioethics more broadly, the configurations and transformations associated with ubiquitous computing have been relatively under-explored. The question of how to think ethically about new healthcare diagnostic tools and computer-enabled accumulation and analysis of large sets of biomedical data is not totally new to bioethics. However, the question has only been systematically problematized with regard to human genetics and genetics-based medicine. The ways in which bioethicists have confronted questions of new and

emerging technologies, such as genetic tests and their associated issues of data ownership, access, and control in connection to genomics, have, in turn, been shaped by the fact that genetic testing and (to a lesser extent) genetic sequencing have been largely integrated into the existing health system or regime. Although the introduction of genetic medicine has had significant implications in a few areas of medicine, such as breast cancer diagnosis, genetic medicine has not yet had a transformative impact on the overarching organization and practice of medicine, though these transformative effects have long been anticipated and contemplated. Moreover, ethical debates about genetic medicine have been largely circumscribed by previous bioethical disputes concerning the autonomy of the individual person and the protection of vulnerable populations [1, 2]. These previous bioethical considerations have been recast in light of the particular contexts within which the ethics of genetic medicine were first articulated: most especially the entanglement of health insurance and employment in the United States and the risks that the genetic screening of individuals might lead to discrimination in employment based on the possible future costs of health care revealed by personal genomics [15]. Hence, bioethics has considered developments at the interface of data collection and new computational powers primarily as problems of privacy, discrimination, and the need for protections against exclusion from health care based on biological identity.

This framing of the ethics of health information has been dominated by two working assumptions that have shaped the ways in which problems and potential solutions have been understood. If adjusted, these assumptions may still prove pertinent to the ethics of biomedicine and new computational technologies. The first assumption is that the potentials and dangers of genetic medicine adhere in the extent to which genomes are taken as a stand-in for an individual's biological individuality. Thus, the sequencing, archiving, and analysis of whole genomes bears deeply on questions of biomedical personhood (and personhood more broadly). The second working assumption is that genetic medicine will be as powerful for intervening in the health of populations, and therefore public health, as it will for intervening in the health of individual persons. Genetic material, after all, is often collected, stored, and analyzed through the matrix of biological interconnectedness. The medical worth of information about an individual's genetic makeup depends largely on the ability to connect and compare that information with the genetic data of much broader populations. Information about the biological identity of the individual and the population are thus imagined and managed as being deeply interconnected.

Bioethics discussions about computer-facilitated collection and analysis of biological data have been dominated by two interrelated concerns. The first is autonomy. In a world of Internet-embedded technologies, individuals must be able to retain control over their own genomic information, including control over who gets to see and analyze it, and under what circumstances [16]. Whether this can be accomplished has remained an open question that becomes particularly problematic in the context of ubiquitous computing, as the volumes of data increase alongside an intrinsic expectation of data sharing. The second concern follows from the first. Given that the control of access to one's genomic data will likely be limited, individuals will need to be protected by privacy and anti-discrimination laws that keep

employers and others from discriminating against them on the basis of information about their biology and biological futures [17]. Given the significance of genetic data with regard to one's real and imagined biological self, as well as psychological and social identity, privacy and discrimination protections are particularly significant.

While considerations of autonomy, privacy, and discrimination remain significant as we shift from thinking about genetics to ubiquitous computing, the ethical challenges of the latter extend well beyond these traditional concerns. The introduction of ubiquitous computing represents a moment of widespread potential transformation of aspects of medicine and health: the day-to-day practice and organization of health institutions; the financial structures of these institutions and the mix of industries involved in the health system; the experiences of individual patients; and the nature and character of medical knowledge. In the rest of the chapter, we discuss three sets of new ethical challenges confronting the healthcare industry, its technology partners, and society more broadly as they engage in this transformation. The first—building on existing bioethical considerations of privacy, autonomy, and discrimination—examines the ethics of *the relationship between the individual and the health system* in the context of dramatic changes in health systems. The second examines ethical quandaries that arise in *the day-to-day practice of medicine* as that practice is reconfigured through the introduction of ubiquitous computing. Finally, the third looks at the broader ethics of *public health, public values, and health technology regulation* in the context of broad shifts in health practice and health system organization.

18.4.1 Transforming the Health System

Fundamental change in the health system is a key element in all three visions of the future of ubiquitous computing in health care. Health system changes drive key actors in the field, whether those actors are information technology companies seeking to disrupt[5] the health industry and open new digital markets by promoting individual patient empowerment, biomedical research institutions looking to transform medical knowledge, or healthcare institutions looking to leverage new technologies to use enhanced medical surveillance to better promote the health of their patients. It is useful to begin our ethical reflections with a better understanding of the nature of this change.

Information is at the center of ubiquitous computing in health care. For many, information is seen as neutral, and more information is seen as inevitably beneficial. However, as the earlier case study of remote interrogation of heart devices demonstrates, the kinds of information generated by personal health surveillance technologies are neither neutral nor do they straightforwardly enhance care. Rather, new information perturbs established notions of what needs to be known, by whom, and under what conditions in order to provide good medical care. In routine medical care, information informs action in relatively well-specified ways. Information is taken as warranting action (or inaction) depending on existing knowledge about health and disease and on

[5] Disruption is a commonly used term in Silicon Valley and in discussions of innovation that denotes the transformative impact of new technologies [18].

how that knowledge has been translated into clinical routines and standards of care. Knowledge of disease also defines the parameters of what needs to be known about a particular patient, and thus the routines of eliciting information that are embedded in clinical practices, diagnostics, and the systems of accounting and reimbursement that distinguish between what is medically necessary and what is not [19].

Each of the three visions of ubiquitous computing in health care described before rests on a variant of the promise that new information from individual health devices will improve health outcomes. Yet each also argues that the benefits of these devices will only be realized by their effective integration into a robust ecosystem of dense, interconnected computational technologies comprising systems of data collection, aggregation, synthesis, and analysis, and feedback of the resulting insights throughout medical care. Just as Apple offers HealthKit and Medtronic offers CareLink, Qualcomm Life offers a platform for linking multiple health devices ranging from active information gathering tools (e.g., weight scales and blood pressure monitors) to passive monitoring devices (e.g., blood glucose meters, inhalers, and activity monitors). The integrated platform is thereby "liberating biometric data so that it becomes ubiquitous across the continuum of care" [20]. Echoing the views of Eric Topol, the IOM's Precision Medicine committee, and a swelling group of thought leaders in the medical community, Qualcomm Life equates "liberated data" with medical benefit. The company is "mobilizing health care to improve lives and advance the capabilities of medical devices" [21]. In addition to device integration platforms, healthcare and biomedical research institutions are increasingly investing in a wide variety of technologies for health data collection, synthesis, and analysis, including genetic screening programs and population genomics databases, diagnostic tools that sample extensive biomarkers, and biobanks for tissue collection and storage.

The proponents of these technologies imagine and are attempting to construct a new infrastructure for health—a health data highway—that differs radically from prior health systems. This health data highway is held together by the notion that ubiquitous and extensive data-gathering is essential for medicine: for new medical knowledge, care of the individual patient, and public health—even though the precise forms, aims and justifications behind specific aspects of that data-gathering are presently ill defined. What traffic will move on this new information highway cannot be known in advance, but traffic will inevitably be directed and constrained by the commitments that are made in shaping the highway infrastructure. In this sense, innovation in health information technologies might be understood as more analogous to urban planning than to product design. The most important ethical issues apply not merely to the development or application of specific devices but also to the ways in which this systemic infrastructure aspires to and ultimately does reshape how people live, their health outcomes, the work, function, financing, and norms of the healthcare system, and broader patterns of public health.

18.4.2 The Individual and the Health (Data) System

The first set of ethical challenges we consider arises precisely because changes in the health system fostered by the rise of new data technologies (e.g., sensing devices, databases, and big data analytics) impact the relationship between the individual

patient and the health system. Simply, the new regimes being constructed to serve ubiquitous computing in health care require exposing the patient's body—and the body's relationship to other facets of living, such as eating habits, chemical exposures, and exercise routines—to constant, thoroughgoing surveillance. The more thoroughly the body and its experiences and environments are known—in any of the three visions described before—the richer the information available to researchers, health practitioners, and the individual him/herself to create the knowledge necessary to improve health.

Informed by decades of social movements for patient empowerment, from the 1970s women's health movement and 1980s AIDS activism to millennial patient advocacy groups like 23andMe, the ethics of health data has focused on notions of autonomy that presuppose that patients have full control of and access to the data and knowledge necessary to make robust, informed health decisions for themselves, including especially their own health data [22, 23]. The *knowledge-empowered patient* vision of ubiquitous health computing draws heavily on this notion of patient autonomy. Yet the technological systems being constructed in the service of this vision—as well as the other two visions we describe—depend on the free flow of so-called liberated data and information from user-centric devices to a wide variety of other data partners. In this sense, ubiquitous computing in health care follows in the footsteps of other information technology domains, such as Facebook and other social sharing sites, where privacy is articulated as a powerful framing device but is significantly undermined by the system's organization to encourage individual users to share as much as information as conceivably possible. The more the data, the more valuable the computational analyses that can be derived from it.

Moreover, as in the Campos case discussed before, there is no guarantee within emerging health data systems that the individual patient will have access to the data generated by health information devices or, even if they do, to the analytic tools and capabilities—or the reams of other patients' data—necessary to make sense of that data effectively. The question of what forms of health-related information individuals should have access to, and how that information should be curated to ensure a reasonable, informed response, is a longstanding problem in medicine. For instance, evaluating and certifying information is the oldest and most foundational responsibility of the US FDA. The FDA regulatory model takes product claims (and the evidence upon which they are based) as the primary object of regulation and, as a result, has engendered a health economy in which certain forms of information are a critical capital resource. Control over what is known and how it is known can be fundamental to the success or failure of firms [24]. Ubiquitous medical computing promises to radically increase the scale and the ease of access to certain forms of information that could enhance the ability of companies and healthcare providers to deliver services and acquire competitive advantage. Configurations of data access, control and ownership will thus empower some actors (and not others) to shape whether and how information transforms medicine.

As the evolution of ubiquitous computing technologies bumps up against conventional notions of autonomy and data privacy in the health industry, considerations of ethics must expand to understand the possibilities that arise in contexts of widespread transfer, exchange, recombination, analysis, and synthesis of health data among

multitudes of patients. We suggest four major dimensions of this ethical challenge. First, as we have observed, patients must expose private and even intimate aspects of their lives and bodies to surveillance and acquiesce to the circulation of the resulting data. Moreover, patients will be encouraged or even required to do so by the emerging health data system. As a knowledge-making enterprise, precision medicine is not a neutral cartographical project, but requires first that the physical, social, and legal landscape of the individual's body be colonized by surveillance technologies, in effect creating new rights of way into previously closed territories of personal life and making the resulting information available to numerous entities across the health industry. We saw this earlier, for example, in the case study, with the patient whose pacemaker data refuted his self-reported exercise regimen.

Second, as a result of this surveillance and exchange, people will be subject to new vulnerabilities. Comparable ethical problems have been deliberated around considerations of security following in the wake of the Snowden affair of 2013. The evidence that the US and other governments have exploited personal data collected via the Internet and communications companies has illuminated a matrix of vulnerability associated with ubiquitous computing. This matrix is generated by the interplay of peoples' extensive use of digital media with the ways in which their habits of data use generate a vast store of user-specific data that are subsequently exploitable—both technically and financially—by the companies that mediate, structure, and strive to reinforce those habits.

The debate that has taken place concerning this matrix of vulnerability has been articulated in a vernacular of privacy versus security. In the United States, the debate is framed as civil liberties versus government intervention. These familiar framings of the ethical challenges of ubiquitous computing are warranted; they illustrate how novel technological developments can be situated within a longer history of legal and political disputes. But these framings draw attention away from another basic fact of the current situation. The technological transformation of peoples' daily habits entails a corollary transformation and delimitation of their identities. Increasingly, people's understanding of themselves and their relations with others are digitally mediated. The infrastructures of this digital mediation are usually offered to the individual either for free or at a very low cost. In trade, the companies who provide and maintain these digital infrastructures generate profits through the surveillance, collection, analysis, sale, and exploitation of peoples' thoughts, feelings, behaviors, and relationships. This habituation of people to digitally transformed identity and commercial exploitation certainly bears on questions of the public and the private. But privacy in this case might be considered more as something an individual claims in response to threatened exploitation according to established social relationships and norms of the digital medical environment, and less as a sufficient characterization of the individual's experience in the digital medical environment itself.

A more adequate characterization of digital privacy might be captured by two classic terms from the analysis of power: exploitation and subjection. Exploitation, in that people are trading their intimacies and relationships for free access to the capacities of ubiquitous computing. Subjection, in that the price to be paid for this freedom is a willingness to live a life under constant surveillance, where results of

surveillance are catalogued and used for a variety of social, economic, and political purposes [25]. The concepts of exploitation and subjection shift attention from an abstract norm like privacy to the particular configuration of social relationships that are insipient in specific technological or economic regimes. Exploitation and subjection demand analysis that acknowledges the ways in which particular technologies and associated practices are embedded in a larger system. This system must be understood as not merely technological, but also economic, social, and normative.

Consideration of exploitation and subjection introduces a third term: *alienation*. In common English usage, alienation entails the process of isolating people from a common good. Classically, in political philosophy, the term alienation refers to the separation of people from aspects of their own nature as a consequence of living in a stratified society. In the case of the effects generated by the diffuse surveillance and data exploitation characteristic of ubiquitous computing, these common and classical understandings can be fruitfully combined. As vital areas of our lives like health become increasingly mediated by and through digital forms, the actors who own and control those digital forms are positioned to be able—technically and legally—to alienate us from them.

Alienation applies at multiple scales, most obviously in the loss of control over the uses of data derived from one's body, but also in the reconfiguration of one's internal moral relationship in practices of self-care. As the body is rendered more open to medical scrutiny, there is a sense in which the patient is subject to alienation from his/her own health. A program of bioinformatics that constantly interrogates the body to elicit markers of risk has the effect of demanding—and potentially coercing—a kind of personal hygiene of risk-mitigation. Indeed, hunting for and responding to pre-disease is a primary aspiration of the project of digital medicine. These two forms of alienation are linked and mutually reinforcing. Proponents of this regime of surveillance see it also as a platform for developing the therapeutic interventions that will become obligatory for the patient living under surveillance, thereby generating enormous wealth for the firms that capitalize upon it [6]. People's data become a form of capital for biomedical innovation, even as the health system demands that people comport themselves in particular ways in response to new health information and the risks it illuminates. Thus, to the extent that a patient's subjectivity becomes enmeshed, through daily practices, with commercially driven ubiquitous computing, the patient also becomes vulnerable to the exercise of power by those who own their data content, which is made up of the stuff of the patient's everyday life.

These existing vulnerabilities are poised to be intensified by the relationship between ubiquitous computing and health care. The aggregation and exploitation of data in the case of health care consists not only of seemingly incidental information such as one's search history, social media feeds, and purchasing habits but also of digital maps of our bodies, our behaviors, and our health, drawn and redrawn in real time. Indeed, one of the particularly worrisome elements of increased health surveillance is that some of the worst ethical failures of modern medicine have occurred precisely in contexts of heavy medical surveillance, especially when patients and their caregivers were unaware of the full extent of both the surveillance and what was being done with the resulting data [26]. This led precisely to the ethical standards that

have evolved around autonomy and privacy within the current bioethics consensus but which will be increasingly difficult to maintain in their current form in relation to ubiquitous computing.

Third, beyond questions of vulnerability, the relationship between biological data and patient identity is significant in its own right. When bioethicists first began to think seriously about the ethics of human genomics, their concerns were rightly primed by the fact that the genome is a vital aspect of an individual's as well as a population's biological identity. One rebuttal of these concerns argued that that we are more than our genomes and that it was important not to confuse one's personal genetic sequence with one's person. But what such rebuttals failed to take seriously enough is the fact that the genome is indeed fundamental to one's individual biology and to one's biological relations across time and space. This is not because our genome determines everything about our lives, but rather because there is very little in our lives, biologically speaking, that does not involve our genomes in some deep fashion. Similarly, today it is vital to take seriously the fact that the biological and other health-related data accumulated through ubiquitous computing generate online profiles of our biological identities. Just as research has shown that patients reconfigure their identities around genetic testing data, it seems certain that they will also do so around other forms of health information generated within a health system reconfigured around ubiquitous data. Indeed, whereas the fact that the meaning of genetic data is often ambiguous to some degree limits the extent of such reconfigurations. The immediate connection of ubiquitous data to everyday life suggest that its effects on identity may be still more profound than those of genomics. These digital residuals of our biomedical bodies are certainly not identical with our biological bodies. But neither are they alien to our bodies. It would be unwise to remain indifferent to the circumstances under which data about our bodies are collected, aggregated, analyzed, and exploited. At present, leading figures in biology and biomedicine are promising that a world in which the differences are collapsed between biology, the digital representations of biology, and the material recapitulation of those digital representations is just around the corner. These promises are surely overstated. Nonetheless, they signal that we cannot remain indifferent to a biomedical reality in which the real-time data that are being accumulated and archived about our bodies are bound to be alienated from us and commercially exploited. Indeed, it is worth remembering that current rules governing personal medical information under the HIPAA were explicitly constructed to permit alienation of individual information by "promot[ing] the use of electronic health information in the [health care] industry" (27; cf. 28).

Finally, fourth, it is also crucial to ethically consider the ways in which new data technologies create and reinforce evolving tendencies within the health system in which the individual patient bears the burdens of responsibility to secure, purchase, and navigate healthcare systems. In recent years, the health system has increasingly reoriented around the idea that the patient constitutes an individual medical consumer who is responsible for knowing and acting in a way that secures his/her own health as maximally as possible. The tendency to shift the burden of obesity, for example, onto personal eating habits, as opposed to emphasizing the structural features of the

food system that make eating healthily extremely difficult for many people, reflects this conceptualization of the responsible individual. We see here how healthcare regimes simultaneously operate under the sign of "personal choice" and "individual empowerment," while creating structural conditions that cause significant numbers of individual consumers of healthcare to become overburdened and excluded.

The hypothesis that ubiquitous computing amplifies this dilemma begins with the observation that, in the name of empowering the individual consumer, a glut of data about one's individual body and the relation of one's body to norms established through population data are being generated at a level of detail, pace, volume, and diversity that outstrips the ability of experts (to say nothing of individual consumers) to define and refine regimes of intervention based on these data. This asymmetry between data and intervention generates a situation in which the demands placed upon the patient-as-consumer to make informed choices about what to do and when to do it are likely to intensify. Such demands are ethically troubling because, despite the fact that this data-rich situation is advanced as an example of the way in which personalized medicine increases the individual's ability to manage him/her own medical destiny, it is likely to produce the same mixture of anxiety and deferral that has been catalogued in other contemporary cases of data overload. It produces anxiety in the feeling that one simultaneously has too much and too little information. It produces deferral in that it reinscribes a dynamic wherein ostensibly empowered users default to the authority and courses of action recommended by the doctors, technicians, or algorithm-driven devices that mediate and communicate the data in the first place.

In short, a key ethical question for ubiquitous computing in health care is how burdens of responsibility will be designed into emerging data systems with regard to two questions: Who is responsible for what kinds of knowledge and expertise, and who is responsible for ensuring that those responsible for knowing have the necessary skills and capabilities to carry out their tasks? When Hugo Campos received the spreadsheet with his device data, who was responsible for interpreting the raw data for him? This question poses a quandary because, on the one hand, any imposition of wholesale responsibility for understanding and proper application of the massive data generated by new health information devices without radical rethinking of basic health education is likely to founder on the inability to patients to live up to that responsibility. At the same time, any allocation of this same responsibility toward others is likely to lead to significant reversals in the flow of power in healthcare settings away from patients and towards providers, insurers, data analytics companies, and others.

18.4.3 Day-to-Day Ethics in Healthcare Practice

The question of responsibility for health data and informed decision-making is a macro-scale question for the health system. At the same time, responsibility is also a micro-level ethical question that arises routinely in day-to-day practice. Moreover, as the case study of medical devices in cardiac care illustrated, it is a question whose assumed-to-be-normal parameters in healthcare settings are significantly perturbed through the introduction of new forms of information devices and their associated

data technologies. These perturbations of day-to-day medical practice and the associated ethical challenges they pose are the second set of ethical dimensions of ubiquitous computing in healthcare that we wish to discuss.

Improving lives by harvesting the (hypothesized) benefits of ubiquitous so-called liberated data first requires that new technological infrastructures be built and that individuals and healthcare institutions incorporate these technologies into their lives and operations, even while their benefits remain ambiguous or underdeveloped. Even in the case of a technology as well established as heart rhythm monitors, we have seen how the mere fact of new streams of information perturbs established routines of clinical care, requiring a period of significant adjustment. Central to this readjustment is the adjudication and negotiation of responsibility, first and foremost between healthcare institutions and patients but also more broadly within the medical profession, between the healthcare industry and regulatory agencies, and potentially involving the courts or legislatures in reconfiguring responsibilities. While healthcare providers in a given clinic may debate among themselves how frequently to remotely monitor wireless implanted devices, what triggers to use, and how to intervene in a person's life and health on the basis of that information, they do so in the full knowledge that (i) patients will likely second-guess their judgments and, in the context of a litigious society like the United States, sue if they dislike the outcomes; (ii) the medical profession itself will debate what constitutes appropriate standards of care; and (iii) lots of people and institutions will be scrutinizing providers' practices and medicine's internal debates.

In the cardiac device case, clinicians had to establish new responsibilities and workflows in order to accomplish near-instantaneous monitoring of continuously collected cardiac device data. These practice changes were designed to afford optimal possible outcomes for patients, as well as to avoid the perception—or possible legal accusation—of negligence in the event of a bad outcome. In the absence of a firm set of guidelines for remote monitoring practices, the cardiology practice felt compelled to demonstrate constant oversight of the new flows of continuous device data.

This is the case for a technology whose utility is well established and for which the relevant forms and uses of diagnostic information are putatively already clear. For devices whose utility is less clear, and which generate information for which the appropriate clinical (or personal behavioral) response is ambiguous, guidelines will be much more significant. Furthermore, whereas heart rhythm data is piggybacking on an already established and medically well-justified integration of a technology into a body, the therapeutic rationale for employing other biometric technologies may be less well defined.

Therefore, consideration of the ways benefit is measured, to whom benefits flow, and who has the responsibility for determining benefit and acting upon it become especially significant in the reconfiguration of medical practice around new devices and information infrastructures. For example, is a monitoring device whose sole purpose is to reduce costs or streamline clinical practice a benefit to the patient? After all, some things are lost in displacing an attentive human being with a machine (even as other things may be gained). In another example, is a suite of devices that reveal correlations between physiological and behavioral states that become the basis

for developing novel, proprietary therapeutics a benefit? To whom? Where streams of device-driven data are used to advance knowledge or innovation and not merely for individual care, there is an inevitable disjunction between patient-focused and health-system focused rationales for integrating these devices into patients' lives.

Within these reconfigurations, it seems obvious to suggest that new health information devices will unequivocally place new responsibilities on physicians and patients. For instance, in 2013 the American College of Medical Genetics (ACMG) issued a policy statement declaring that if a patient undergoes genome sequencing and an incidental finding is made of certain variants of known medical significance, the doctor must inform the patient, even if the patient has asked not to be informed [29]. Though the policy received significant criticism and the ACMG has since modified it to allow patients to refuse to be informed, the example illustrates how technologies that seem purely enlightening in the abstract can, in the context of existing high-stakes configurations of responsibility and liability, come to impose unexpected roles and responsibilities upon their users [30]. These are not unintended consequences of technology but rather are expressions of the challenges of integrating new technologies into established social and normative arrangements. That this controversial moment could erupt in the domain of human genetics, where decades of thinking has already been devoted to anticipating ethical questions, suggests that much more attention must be given to the social and normative environments into which digital health technologies are integrated, and that these considerations should inform the very design and deployment of the technologies themselves.

18.4.4 Public Health and Public Goods

Our final set of ethical reflections concerns questions of political economy—large-scale distributions of wealth and power—associated with changes in health systems that derive from ubiquitous computing in health care. We raise these considerations not only because distributions of wealth and power are central to questions of ethics and justice, but also because they connect fundamentally to problems of public health, in the sense of the ability of the healthcare system to deliver health to publics broadly, as a public good. Recall, for example, that the FDA was established and continues to operate today largely to police the introduction of new technologies into healthcare markets. The FDA has this power precisely because medical technologies have the potential to disrupt as well as to enhance public health and because a proper securing of health advantageous technologies (whether through ensuring the veracity of the health benefit claims made by proponents of new technologies or, later, ensuring the safety and efficacy of new drugs and devices) would significantly enhance the public good and disrupt the inappropriate ability of those with wealth and power to exploit individuals through control over relevant information [31].

A major driver of the proliferation of the so-called Internet of Things has been the coordinated transformation of individuals' day-to-day habits with regard to how they use personal digital devices and services (e.g., smartphones, tablets, and social media). The users of these devices have articulated the stuff of their everyday lives into and onto digital platforms. The extent of this articulation can be seen in the ways

in which individuals now increasingly embed their relationships into digital media as part of their daily routines and in the way in which the use of handheld devices has transformed norms of bodily comportment. Ubiquitous computing, as it is experienced by many people today, has transitioned from an interconnected set of tools to a digital milieu through which and in which people are conducting much of their lives. To the extent that health care is integrated into this digital milieu, it seems likely that ubiquitous computing—and the projected hopes and fears attached to it—will become part of a regime of daily expectations.

A pressing concern is the economic and actuarial impact of such shifts in expectations, catalyzed by the inter-articulation of the daily use of digital devices and new regimes of healthcare. The economic impacts of new technologies in medicine are well known. Market-driven proliferation of new technologies exacerbates the overall cost of health care. It is less clear whether these familiar dynamics will be reproduced in the economics of ubiquitous computing. To what extent will widespread investment in the institutionalization of ubiquitous computing increase the proliferation and use of other costly infrastructures? Or, equally troubling, might the ancillary costs of ubiquitous computing in healthcare reproduce the economics of ubiquitous computing in other related domains? For example, will we see a pernicious trade-off with relatively inexpensive access to data and analytics for individual users at the cost of relinquishing ownership and rights to the commercial use of those data and analytics by the enterprises, such as Apple, that provide inexpensive computational services? Based on current trends, this scenario is highly likely.

An actuarial question closely follows. Will the economics of ubiquitous computing intensify existing disparities in health and access to health care? A story of innovation in computing has long been that as technologies become more and more diffuse they become less and less expensive. But it is far from obvious whether this will be true in the case of health care. If the valorization and implementation of ubiquitous computing in medicine does intensify health disparities, it is fair to ask whether and to what extent it should continue to receive support.

Equally significantly, a second actuarial question focuses on what medical austerity will mean for trade-offs between new forms of digital health care and existing care or alternative new approaches. To the extent that new infrastructures require extensive new investments by health institutions and/or monopolize the available human and financial capital for digital health systems, they may force trade-offs that significantly weaken other aspects of healthcare systems, in turn impacting health in unexpected ways. For example, clinicians have devoted time previously spent on in-office patient visits to monitoring remotely transmitted device data. This has, in some cases, resulted in fewer available office appointments and longer waits for patients to schedule a routine checkup with a heart rhythm expert. In other cases, by replacing multiple face-to-face visits per year with telemonitored data exchanges, other physical findings like subtle skin reactions to long-term medications go unrecognized and untreated until permanent damage is done.

These questions raise the issues of regulation. FDA regulation in this domain narrowly focuses on specific devices, not on the larger systems that they contribute to. A device might be approved for a narrow diagnostic application, but it could

generate far more data than are necessary for the approved diagnostic application. Or it might generate data that are necessary to that application but which can also be put to myriad other uses that do not require regulatory approval because they do not directly affect the patient. In such a scenario, data becomes a valuable byproduct of medically sanctioned uses of a technology; capturing the value of this by-product may be part of the very business plan that underwrites the development of the technology in the first place. For the manufacturing firm, the data that the device elicits may be far more valuable than the initial revenue stream from selling the device. Such data might be the basis for amassing the information resources necessary for developing or validating new diagnostics and thus for extending its market share. Therefore, the actual rationale for authorizing the use of a device in medical care may be far narrower than the range of imaginable—and intended—uses. In such an arrangement, patients would pay for the privilege of serving as unwitting experimental subjects in a regime of technology development, even as a firm's monopoly on the resulting body of data is used to create new dependencies.

Myriad genomics offers a good example of precisely this strategy. In the past decade, the company has leveraged its patent on the BRCA genes to build a database of variants that it now treats as a trade secret. Because every woman seeking BRCA testing had to go through Myriad, the company had exclusive access to a huge number of genetic samples. It has used these to develop a multi-tiered diagnostic product that is offered for dramatically more than its cost. Even with the end of its patent monopoly after the US Supreme Court invalidated its key patents, Myriad has maintained this monopoly [32]. Because it has this proprietary database of patient-derived genomic data, Myriad's diagnostic capabilities are superior to competitors. By enjoying a monopoly on a diagnostic technology, Myriad also acquired a monopoly over an information resource that was extracted from patients' bodies. The benefits of this monopoly have made Myriad highly profitable, at patients' expense, and, given the financial barrier to access the test, likely also at the expense of public health [33].

Scenarios like these give further urgency to the question of how benefits will be defined, and by whom? Who will define the needs to which these devices putatively respond and according to what measures? These questions are further complicated by the fact that the project of ubiquitous medical computing is underwritten by a promise of future benefit. The technological infrastructure must exist in order for its anticipated systemic benefits to be realized. Leading figures who argue for committing to this infrastructure often characterize benefits in terms of justice. Leroy Hood predicts that the digitization of health will increase effectiveness, reduce costs, and increase access to include the world's poor [6]. In the Xprize's tricorder competition, contenders are building consumer-focused devices that can monitor and diagnose a wide range of health conditions to bring "healthcare to the palm of your hand" [34]. One explicit aim of the competition is to bring health care to the developing world in the form of digital devices. However, from a public health perspective, in the developing world, a computer that generates more precise diagnoses to those in conditions of profound basic medical need is a rather attenuated imagination of health care.

Taken together, the ethical issues enumerated before reveal the complex stakes of ubiquitous medical computing. They suggest the profound reorientations of medicine

that this technological infrastructure entails and also the ethical stakes of simply imagining and aspiring to a future in which medicine is suffused with bioinformatic-data eliciting devices. In aspiring to this technological future, participants are simultaneously reimagining—and reorienting—social and normative relationships in biomedicine. They are, in effect, committing to the social, legal and ethical adjustments necessary to open rights of way for integrating these technologies into our lives.

These reorientations are significant. Though reorientations may be framed simply as logistical challenges of getting at data that, once liberated will inevitably benefit us all, they have far-reaching effects. For instance, the notion that information inevitably empowers may reframe issues of distributive justice as problems of access to information. Or, the anticipated collective benefits of "big data" may justify sidelining traditional concerns about power relations—often articulated in terms of privacy—in the name of the imperative to advance biomedical progress (e.g., Ref. [35]). In adopting particular technologies in the context of such visions, we are also making commitments to a systemic, infrastructural regime and to the normative commitments that organize that regime, including tacit allocations of rights, roles, and responsibilities.

When innovation is assessed narrowly at the level of individual technologies, the systemic commitments implicit in the emergent socio-technical regime, including ethical commitments, remain largely invisible. The promises of—and demands for—a transformed social world are not currently subject to regulatory validation. Thus it is an ethical imperative that we collectively recognize, imagine, and govern the formation of that regime. If we circumscribe questions of public protection and welfare to the process of regulating only each brick and not the edifice that they together form, the barriers and exclusions that the edifice imposes will be known to us only after it has already been erected.

18.5 CONCLUSIONS: THE NEED FOR SOCIO-TECHNICAL DESIGN

In the nineteenth century, changing notions of how disease should be interrogated and how medical knowledge should be advanced led doctors to establish a new kind of institution—the clinic—whose purpose was to facilitate and render routine the surveillance of patients who needed constant monitoring [36]. Today, intensive and neonatal care units, nursing homes, and a much-reduced number of inpatient psychiatric wards are the remnants of this once thriving form of institution. Their decline has resulted from several factors, including their extremely high costs but also because some of history's most egregious exercises of medical authority and lapses of medical ethics occurred in these settings. In part, clinics facilitated ethical problems by isolating patients from their families, but underlying power dynamics also contributed. Since the 1970s, in particular, the pendulum of power relations between doctors and their patients has swung dramatically toward patients. Patient health movements—especially among women but also among disease groups—have significantly strengthened the knowledge of patients, emboldening them to claim power over health decisions. So too have regulatory changes, for example, rules at the US

FDA that have allowed advertising and information on drugs to be made available to the public, as well as a variety of strategies for holding the healthcare industry more accountable both for its outcomes and for its ethics.

Smart medical devices and the revolution they promise in bringing ubiquitous computing into healthcare systems play deeply into the rhetoric of patient empowerment, pledging to further put patients in control of their own health information and thus their own health, perhaps. But this will come to pass only insofar as ubiquitous computing is approached as a project of socio-technological system design.

We have identified in this chapter the following four significant challenges to the prospects for empowering patients in the field of smart health information devices:

1. *Knowledge is power, but health information may ultimately not be controlled by patients or may simply empower other players in the healthcare system more effectively than it empowers patients.* Healthcare providers, insurance companies and other payers, and medical device and software companies are likely to find strong reasons to seek out greater health information—if it is available—and to have the institutional tools and knowledge to make use of that information. The trends in medical testing already taking place suggest the scope of the problem, as testing regimes are rapidly growing and under the control of doctors, laboratories, device makers, and insurance companies.
2. *Health information will engender new responsibilities for both patients and healthcare providers, but is likely to be difficult to translate into better health decisions.* Anyone who has ever wandered into the personal health section of their local bookstore will recognize the fallacy of assuming that more information necessarily leads to better outcomes. One of the primary promises of ubiquitous medical computing is to profoundly increase the amount of information that people have about their bodies and their health. The sheer existence of this information is bound to be overwhelming. Furthermore, it is likely to be of varying quality and significance. This poses an obvious practical challenge for the would-be dutiful knowledge-empowered individual. But the challenge extends well beyond the individual to the institutional authorities and experts who adjudicate between good and bad information and declare what information warrants action and under what circumstances. This has been a major concern in genomics, particularly vis-à-vis direct-to-consumer genetic testing. In evaluating these technologies, the FDA was concerned to assess not merely the accuracy of genetic risk analyses offered by companies like 23andMe but also the ways that consumers understood—and acted on—that information. The question remains far from settled, but the key point is that the stakes turn simultaneously on the quality of the information, the ways the average consumer is likely to understand that information, and the normative tension between a consumer's right of access to the information and the state's responsibility to protect the consumer from harm [37]. This problem is likely to be replicated many times over in the domain of ubiquitous health computing.
3. *Accountability for health outcomes remains low and may be even more problematic in a world of ubiquitous health computing.* Except in rare cases,

neither patient, nor provider, nor insurer are either rewarded or held accountable based on individual health outcomes. Nor, at the end of the day, will health information device manufacturers or operators. It would be nice to think that markets are rational and that smart medical devices will be sold based on whether they really help patients improve their health. But that too is a fallacy that falls short of the mark. Historically, as the rise of evidence-based medicine has shown, many medical interventions have not improved outcomes. At the same time, much of the promise of health information devices is predicated on their *future* benefits: benefits that will fully arrive only after they are in widespread use. Hence the market, such as it is, is bound to be driven by expectations rather than outcomes. One key outcome will be the extent to which expectations shape particular regimes of surveillance, data ownership, and control. Such shaping may very well unfold largely independently of specific, measurable effects upon *individual* health outcomes.

4. *Medical austerity may force trade-offs to offset the costs of health information devices and monitoring services by reducing coverage of other medical services and so reducing overall health.* In the end, health care can only consume so much of the overall economy and finite government program resources. Health information devices enter a crowded healthcare market that may struggle to pay for new kinds of services and face difficult choices if the costs of ubiquitous computing rise significantly without offsetting other medical costs. One important concern with health information devices will be the extent to which investment of scarce medical resources is driven by *anticipation* of future cost reductions through device-driven efficiencies or health outcomes. The massive governmental and institutional investments in building electronic medical record infrastructure is a case in point. Thus far, these investments have increased institutional costs but have yet to deliver on promises of increased productivity or reform of clinical practices through health informatic-based clinical effectiveness research. Electronic medical records and health information devices alike may eventually yield meaningful gains. But particularly with such devices, the answer cannot be known in advance, and judgments about just allocation of limited health resources may be made more on confidence in a vision than on the basis of well-established bodies of evidence.

Addressing these challenges will require more than just creating devices that are smart. Patients, providers, healthcare institutions, and regulators will all need to be smart about how they design and inhabit the socio-technical relationships that link health information devices to values, behaviors, sensibilities, relationships, and institutional practices. Smart designs will focus on achieving good health outcomes, but not exclusively. Other outcomes matter, too. Privacy and security in new health information systems will need to improve dramatically over their current counterparts in the health industry, along with new strategies for understanding and managing ownership and control. The long-term allocation and institutionalization of power among patients, providers, entrepreneurs, and governments will matter enormously.

Health justice will require vigilant and ongoing scrutiny of benefits, costs, and risks with the aim of checking the balance of those benefits, costs, and risks across patient groups and other consumers. It will equally require inclusivity in articulating and giving form to the imaginations of needs and goods that will govern judgments of justice and injustice.

Today's healthcare system is one of the most complex socio-technological systems in modern societies, as are the socio-technological systems that constitute the emerging Internet of Things. The adoption of ubiquitous computing means bringing these two systems together. This reassembling of systems is as likely to be disruptive, as it has been in other domains of contemporary life. In complex systems, however, disruptions can have wide and deep consequences that ramify in ways that travel far beyond the initial points of intervention. The transition to ubiquitous computing is likely to be more ethically sound if we can learn to anticipate how disruptions may travel, and design socio-technical systems that are attuned to the new worlds—and the corollary social, political, and economic transformations—that people usher into being as they put new technologies to use in expected and unexpected ways.

REFERENCES

[1] President's Council on the Study of Bioethical Issues (PCSBI). 2012. Privacy and Progress in Whole Genome Sequencing. Washington, DC: US GPO.

[2] Chadwick, R., M. Levitt, and D. Shickle, eds. 2014. The Right to Know and the Right Not to Know: Genetic Privacy and Responsibility. Cambridge, UK: Cambridge University Press.

[3] Topol, E. 2013. The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care. First Trade Paper Edition, Revised and Expanded Edition. New York: Basic Books.

[4] Young, N. 2012. The Virtual Self: How Our Digital Lives Are Altering the World Around Us. Toronto: McLelland and Stewart.

[5] IOM and NRC Committee on a Framework for Development a New Taxonomy of Disease and National Research Council. 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: The National Academies Press. Available at: http://www.nap.edu/openbook.php?record_id=13284, accessed on January 5, 2016.

[6] Hood, L. 2013. "Systems biology and P4 medicine: Past, present, and future." Rambam Maimonides Medical Journal 4 (2): e0012.

[7] Aquilina, O. 2006. A brief history of cardiac pacing. Images in Pediatric Cardiology 8(2): 17–81.

[8] Burri, H. and D. Senouf. 2009. Remote monitoring and follow-up of pacemakers and implantable cardioverter defibrillators. Europace 11(6): 701–709.

[9] Lakshmanadoss, U., A. Shah, and J. P. Daubert. 2011. "Telemonitoring of the Pacemakers," In M. R. Das, Ed. Modern Pacemakers—Present and Future. InTech. Available at: http://www.intechopen.com/books/modern-pacemakers-present-and-future/telemonitoring-of-the-pacemakers, accessed on January 5, 2016.

[10] U.S. Department of Health and Human Services (HHS), Health Resources and Services Administration, National Center for Health Workforce Analysis. Projecting the Supply and Demand for Primary Care Practitioners through 2020. Washington, DC: U.S. Department of Health and Human Services, 2013.

[11] U.S. Department of Health and Human Services (HHS), Health Resources and Services Administration, National Center for Health Workforce Analysis. Projecting the Supply of Non-Primary Care Specialty and Subspecialty Clinicians: 2010–2025. Washington, DC: U.S. Department of Health and Human Services, 2014.

[12] Vinck, I., C. De Laet, S. Stroobandt, and H. Van Brabandt. 2012. "Legal and organizational aspects of remote cardiac monitoring: The example of implantable cardioverter defibrillators." Europace 14: 1230–1235.

[13] Slotwiner, D. et al. "2015 HRS Expert Consensus Statement on Remote Interrogation and Monitoring for Cardiovascular Electronic Implantable Devices." Available at: http://www.hrsonline.org/Practice-Guidance/Clinical-Guidelines-Documents/Expert-Consensus-on-the-Monitoring-of-Cardiovascular-Implantable-Electronic-Devices/2015-Expert-Consensus-Statement-on-Remote-Interrogation-and-Monitoring-for-CIEDs, accessed on February 11, 2016.

[14] Wiltz, C. 2015. "Access to Your Device Data? That'll Be $797. MDDI Medical Device and Diagnostic Industry News Products and Suppliers." Available at: http://www.mddionline.com/blog/devicetalk/access-your-device-data-thatll-be-797, accessed on January 16, 2015.

[15] Quinn, G., A. de Paor, and P. Blanck. 2014. Genetic Discrimination: Transatlantic Perspectives on the Case for a European Level Legal Response. London: Taylor & Francis.

[16] Erlich, Y. and A Narayanan. 2014 "Routes for breaching and protecting genetic privacy." Nature Reviews Genetics 15 (6): 409–421.

[17] Prince, A. E. R., and M. Waterstone. (2014) "The Genetic Information Nondiscrimination Act (GINA) of 2008." In G. Quinn, A. de Paor, & P. Blanck Eds., Genetic Discrimination: Transatlantic Perspectives on the Case for a European Level Legal Response. Oxford: Routledge, pp. 114–127.

[18] Christiansen, C. 1997. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Boston, MA: Harvard University Press.

[19] Rosenberg, C. E. 2002. "The tyranny of diagnosis: Specific entities and individual experience." The Milbank Quarterly 80 (2): 237–260.

[20] "What Is 2net? | Qualcomm Life." 2014. Available at: http://www.qualcommlife.com/wireless-health, accessed on December 6, 2015.

[21] "Company Overview | Qualcomm Life." 2014. Available at: http://www.qualcommlife.com/company-overview, accessed on December 6, 2015.

[22] Epstein, S. 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. Berkeley, CA: University of California Press.

[23] Saha, K. and J. B. Hurlbut. 2011. "Research ethics: Treat donors as partners in biobank research." Nature 478 (7369): 312–313.

[24] Dumit, J. 2012. Drugs for Life: How Pharmaceutical Companies Define Our Health. Durham, NC: Duke University Press.

[25] Niccol, A, et al. (1997). Gattaca (motion picture). Culver City, CA: Columbia Pictures.

[26] Hornblum, A. M., J. L. Newman, and G. J. Dober. 2013. Against Their Will: The Secret History of Medical Experimentation on Children in Cold War America. New York: Macmillan.

[27] U.S. Department of Health and Human Services. 2007. “HIPAA Security Series.” Available at: http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/security101.pdf, accessed on January 5, 2016.

[28] Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. 2009. “Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research.” In S. J. Nass, L. A. Levit, and L. O. Gostin, Eds. The National Academies Collection: Reports Funded by National Institutes of Health. Washington, DC: National Academies Press. Available at: http://www.ncbi.nlm.nih.gov/books/NBK9578/, accessed on January 5, 2016.

[29] Green, R. C., J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, et al. 2013. “ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.” Genetics in Medicine 15 (7): 565–574.

[30] ACMG Board of Directors. 2015. “ACMG policy statement: Updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing.” Genetics in Medicine 17 (1): 68–69.

[31] Carpenter, D. 2010. Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA (Katznelson, I., E. Schickler, Shefter, M., and T. Skocpol), Princeton Studies in American Politics: Historical, International, and Comparative Perspectives. 1st ed. Princeton, NJ: Princeton University Press.

[32] Cook-Deegan, R., J. M. Conley, J. P. Evans, and D. Vorhaus. 2013. “The next controversy in genetic testing: Clinical data as trade secrets?” European Journal of Human Genetics 21 (6): 585–588.

[33] Parthasarathy, S. (2007) Building Genetic Medicine. Cambridge, MA: MIT Press.

[34] “Qualcomm Tricorder XPRIZE.” 2014. Qualcomm Tricorder XPRIZE. Available at: http://tricorder.xprize.org/, accessed on December 9, 2015.

[35] Kohane, I. S. and R. B. Altman. 2005. “Health-information altruists—A potentially critical resource.” The New England Journal of Medicine 353 (19): 2074–2077.

[36] Foucault, M. (1973) The Birth of the Clinic. United Kingdom: Tavistock Publications.

[37] Green, R. C. and N. A. Farahany. 2014. “Regulation: The FDA is overcautious on consumer genomics.” Nature 505 (7483): 286–287.

19

PRIVACY PROTECTION OF ELECTRONIC HEALTHCARE RECORDS IN e-HEALTHCARE SYSTEMS

FREDRICK JAPHET MTENZI

School of Computing, College of Sciences and Health, Dublin Institute of Technology, Dublin, Ireland

19.1 INTRODUCTION

The use of information and communications technology (ICT) in health care has led to significant improvements in care quality and costs reduction. In this section, a detailed description and analysis of EHRs is given. The advantages and disadvantages of this medical records digitization exercise are discussed. It is shown that the huge amount of sensitive personal information that is collected, processed, transmitted, shared, and stored is the necessary evil for e-Healthcare Systems to provide the required services.

19.1.1 Introduction to EHRs and e-Healthcare Systems

The healthcare spending accounts for a sizable portion of most countries gross domestic product (GDP), for example, according to the Organisation for Economic Co-operation and Development (OECD) the United States spends more than most developed countries [1] and least developed countries spend less. This trend of spending is set to grow with modernization of health care especially the use of

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

information and communications technology (ICT), which studies referenced in this chapter show improves the quality of care and reduces costs when fully adopted. However, the proliferation of use of ICT in health care opens up security and privacy concerns and challenges not found in paper-based healthcare systems.

It is currently accepted that the quality of healthcare services such as diagnostic, continual data collection, and treatment depends heavily on the utilization of ICT. Digitized medical records called electronic healthcare records (EHRs) [2] are becoming a common way of storing, processing, and communicating health information. The EHRs were originally being stored locally within a certain infrastructure such as medical institution and accessible only from within the infrastructure. However, now the EHRs are integrated and available across several medical institutions and provide a much better complete picture of patient health information. This information is accessible even on consumer devices such as smartphones and tablets. In a few situations, patients are able to manage their medical records in what is called personal health record (PHR). While this access convenience is a welcome addition to healthcare provision, the security and privacy challenges are insurmountable.

The technological developments in the deployment of EHR in practice are impressive, but the acceptance among the population is crowded with security and privacy concerns. A lot of research efforts have been devoted in ensuring that the security and privacy challenges emanating from using either EHR or PHR are addressed or mitigated. For example, the three fundamental security goals of confidentiality, integrity, and availability are essential in e-Healthcare Systems. As EHRs contain information that is considered to be highly sensitive, there is a strong need for confidentiality.

Threats to patient privacy and information security can be categorized into the following broad areas:

1. Organizational threats that arise from inappropriate access of patient data by either internal agents abusing their privileges or/and external agents exploiting vulnerabilities of healthcare information systems.
2. Systemic threats that arise from an agent in information flow chain exploiting the disclosed data beyond its intended use.
3. Threats arising from extensive exploitation of patient data by using techniques such as big data analytics for data collected from mobile and medical devices and inferences made from fragmented personal data found in systems and the Internet. The challenges of patient privacy in the personalized medical era will be formidable.

Privacy is an underlying governing principle of the patient–physician relationship for the effective delivery of healthcare. Patients are required to share information with their physicians to facilitate correct diagnosis and determination of treatment. Over time, a patient's medical record accumulates significant personal information including identification, history of medical diagnosis, digital renderings of medical images, treatment received, medication history, dietary habits, genetic information, psychological profiles, employment, income, and physicians' subjective assessment of personality and mental state, among others.

Many businesses that handle health information are subject to virtually no oversight, and the main source of policy regarding the use of health information comes in the form of self-regulation by the parties that stand to benefit the most from capturing user's health data. However, self-regulation has proven wholly insufficient. Self-regulation is even worse when you consider this scenario: a stolen credit card may be reissued, but if health-related information were leaked online, it could have a devastating impact on millions of people. Merely storing personally identifiable information on health conditions raises the potential for loss, theft, and abuse.

In this chapter, we use the following definitions: healthcare professionals include doctors, dentists, pharmacists, osteopaths, and chiropractors. Personal data refer to the data that are about an individual who can (reasonably) be identified or identifiable. De-identification is the process of removing (or modifying) identifiers from the personal data so identification is not reasonably possible. Pseudonymization is the step where a pseudonym or code is added to this de-identified data. Proportional or reasonable anonymity applies to de-identified/pseudonymized data that cannot be reasonably used to identify specific individuals [3].

19.1.2 The State of the Art of Security and Privacy EHR in e-Healthcare Systems

We are now moving from the era where healthcare systems were single, isolated units to large, diverse, interoperable, integrated systems. These systems use ICT extensively to ensure continuity of care. We present three health-related usages: demonstrating the trends in medical information technologies and the security and privacy challenges they pose are discussed in Section 19.2.

Health sensing: There's been a sharp increase in the quantity and variety of consumer devices and medical sensors that capture some aspect of physiological, cognitive, and physical human health. The implementation of these technologies empowers the end users (e.g., chronic patients) by providing means to monitor and record the status continually and, if the need arises, seek remote assistance [4].

Big data analysis in health care: With the increasing digitization of health care, a large amount of healthcare data have been accumulated and the size is increasing at an unprecedented rate. Discovering the deep knowledge and values from the big healthcare data is the key to deliver the best evidence-based, patient-centric, and accountable care [5].

Cloud computing in health care: With healthcare providers looking at solutions to lower the operating costs, emerging technologies such as cloud computing can provide an ideal platform to achieve highly efficient use of computing resources, simplify management, and improve services. Cloud computing can support the analysis of the big data. There is no doubt that the adoption of these innovative technologies in medical fields can create significant opportunities. Nevertheless, many challenges still need to be addressed in order to achieve truly enhanced healthcare services, especially security and privacy [6]. Accountability and auditing when medical records are accessed and manipulated must be provided by the e-Healthcare Systems. Therefore, there must be a system of checks and balances that is implemented and followed religiously but which continues to allow the data access necessary to perform a task.

19.1.3 Advantages and Disadvantages of Deploying and Using EHR

This chapter discusses and proposes solutions for privacy protection of EHR in e-Healthcare Systems in the era of nano-based technologies. Purely technical solutions are not enough, as they require a relatively high level of knowledge and technical expertise on the part of the patient. By addressing policy and legislation issues directly, rather than combating sophisticated protection techniques, we may produce reasonable and acceptable solutions that outlast today's technology circle.

There are a lot of advantages of using EHR in health care such as significant reduction in healthcare costs, reduction in medical errors, and improved quality of care. In order for these advantages to be realized, an e-Healthcare System must be connected and integrated to provide anytime, anywhere healthcare information and decision support via a comprehensive knowledge-base of interoperable systems. Other potential secondary benefits of EHR adoption is the reduction of care variability by using data to define and disseminate best practices, therefore helping to deliver more effective care to a broader patient base [7]. In addition, consumer and patient interfaces with EHR systems may yield valuable data that might provide additional benefits such as "determining provider (hospital and physician) performance outcomes, monitoring chronic diseases, monitoring medication adherence, promoting safety metrics, determining patient satisfaction, promoting more informed clinical decisions, and improving patient–physician communication tracking" [8].

These and other challenges lead to unintended consequences of advantages of using EHR, such as the ones pointed out by Bernat [9], which author terms *harmful shortcuts* (i.e., copying and pasting of data obtained from other physicians, authorship ambiguities, inadequate discharge summaries, and impaired physicians–patient communication).

The introduction and use of EHR can have a profound effect on medical malpractice liability. For example, Mangalmurti et al. [10] explored this area and raised several points regarding the new responsibilities medical staff and service providers will have to bear following adoption of e-Healthcare Systems including the documentation of clinical findings, the recording of test results, and the use of clinical decision support systems. Should errors arise during the use of the system how will the players be held accountable? The temptation to copy and paste past entries instead of retaking medical histories still exists, and may result in a failure to incorporate new information and the perpetuation of the past mistakes.

The use of ICT and medical devices in e-Healthcare Systems generates huge amount of data that are easily available to healthcare professionals. There is a school of thought emerging, advancing the argument that this new and easier access to information may also lead to an overload that could cause the healthcare professionals to miss important items of information. This has impact on malpractice litigation processes, because the information now available to healthcare professionals is also available to the legal authorities. Integrating clinical systems into practice may redefine the standards of care and provide definite objective answers to previously ambiguous questions, which were caused by incomplete or fuzziness in data.

Kisilev et al. [11] show that it is possible to generate medical reports that are acceptable by health professionals automatically from breast medical images. They further argue that their proposed method should be useful in general doctors' practice, wherein there is a predefined set of medical descriptors to be acquired by a doctor during image investigation. This means automatic information generated from images can enable doctors with little training in reading breast medical images to provide initial informed opinion to patients.

The World Health Organization predicts that chronic diseases will account for almost three-quarters of all deaths worldwide by 2020 [12], so the evolution of M-Health (mobile diagnostics, bio-feedback, and personal monitoring) is set to revolutionize treatment of conditions such as diabetes and high blood pressure. apps designed by medical professionals will provide efficient real-time feedback, tackle chronic conditions at a much earlier stage, and help to improve the lifestyles and life outcomes of communities in the developed and developing world [13].

However, the disadvantages are many such as significant losses of personal information privacy as a result of poorly configured systems, defective safeguards by healthcare providers, or negligent technical system design without satisfactory security safeguards. Other disadvantages include medical identity theft, marketing firms, employer and insurance companies accessing medical data, and sharing of medical information without patient knowledge and consent. Other obstacles limiting the deployment of EHR systems are funding, technology, attitude, and organizational aspects [14]. Overall, more research is needed to determine the best way to implement the system so that it will interfere as little as possible with doctor's daily flow.

19.2 SECURITY AND PRIVACY CONCERNS OF EHR IN e-HEALTHCARE SYSTEMS

The security and privacy challenges of digitized systems have been well researched. This section discusses specific security and privacy concerns as found in the e-Healthcare industry. One characteristic of the e-Healthcare Systems is its generation and use of sensitive patient's data that are extremely difficult to protect.

19.2.1 Security Concerns of EHR in e-Healthcare Systems

Since the beginning of ICT being used in delivery of healthcare, security concerns and challenges have been the major problems. The security concerns and challenges can be looked at through the lenses of the three fundamentals of security: confidentiality, integrity, and availability. According to the ISO EN13606 standard [15], confidentiality refers to the "process that ensures that information is accessible only to those authorised to have access to it." Integrity refers to the duty to ensure that information is accurate and is not modified in an unauthorized fashion. The integrity of health information must therefore be protected to ensure patient safety, and one important component of this protection is that of ensuring that the information's

entire life cycle is fully auditable. Availability refers to the "property of being accessible and useable upon demand by an authorised entity." The availability of health information is also critical to effective healthcare delivery and continuity of care. Health informatics systems must remain operational in the face of natural disasters, system failures, and denial-of-service attacks. Security also involves accountability, which refers to people's right to criticize or ask why something has occurred.

19.2.2 Privacy Concerns of EHR in e-Healthcare Systems

EHR contains some of the most personal sensitive data. Protecting this confidentiality is therefore essential if the privacy of subjects of care is to be maintained. Privacy involves access control versus any party not authorized to access the data, and has been defined as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [14].

19.2.3 Security Threats Classifications Taxonomy for EHR and e-Healthcare Systems

One of the main threats to EHR security is unauthorized access by healthcare staff. Threats from healthcare staff can be divided into two categories: (i) unauthorized access and (ii) lack of user training. Human error is the most challenging issue that needs to be taken into consideration. It may happen in the level of access within the organization with a dramatic effect on the system. The organization could mitigate this threat by conducting proper training and increasing human understanding. Once the security system has been established, an audit function is required. Audit is needed in order for administrators and users to detect any illegal or unauthorized breach.

One of the main privacy threats when sharing EHR data is identity disclosure (also referred to as re-identification), which involves the association of an identified patient with their record in the published data. Identity disclosure may occur even when data are de-identified (i.e., they are devoid of identifying information). This is because publicly available datasets, such as voter registration lists, contain identifying information and can be linked to published datasets, based on potentially identifying information, such as demographics [16], diagnosis codes [17], and lab results [18]. The increased use of mobiles, medical devices, and third-party services such as cloud computing further amplifies the identity disclosure problem.

19.2.4 Privacy Taxonomy for EHR

One of the approaches to ensuring privacy in e-Healthcare Systems is role-based access control (RBAC). A lot of research has been done in RBAC leading to increasingly sophisticated models that provide privacy and/or security [19]. Cryptographic techniques both hardware and software are another cornerstone of ensuring privacy [20]. Privacy protection of medical data that have been encoded as medical codes can be

achieved by using disassociation (an operation that segments records into carefully selected subsets. By doing so, the approach is able to preserve data utility significantly better than existing methods that are based on data generalization or suppression) approaches [21]. However, the development of techniques that preserve medical data privacy is very challenging, due to the complex semantics of medical data and the specific requirements of medical data owners, custodians, and data recipients. This challenge is made even more difficult by the current personalization of medical data generated from mobile and medical devices making most of existing approaches to be insufficient. Text de-identification is one of the most researched directions in the area of medical data privacy. The goal of these methods is to ensure that text data can be shared in a way that does not reveal any health information that must be kept confidential [22].

19.3 PRIVACY LAWS AND REGULATIONS OF EHRs

Privacy laws and regulations of EHRs are playing an important role in the deployment and use of e-Healthcare Systems. They help regulate and ensure that healthcare providers provide services that are fit for purpose. In this section, a detailed discussion of laws, regulations, standards, and policies is given. We clearly show that while there are many laws, regulations, standards, and policies in existence, its landscape is littered with fragmented and old laws that are lagging behind latest technology.

19.3.1 Current Privacy Laws and Regulations of EHRs

While the security and privacy concerns and challenges in e-Healthcare Systems are well documented and understood, how to implement security and privacy in this unique environment is not evident as the balancing of individual and societal interest can be difficult [23]. We have technical mechanisms, methods, and protocols required to provide EHR with protection and privacy. They are developed, tested, and sophisticated. The major challenge is deciding the level and where to place them in the e-Healthcare Systems.

Historically, the European environment for privacy has been uniquely different from that of the United States. The 1950 Council of Europe Convention identified individual privacy as a fundamental value [23]. Article 8 of the European Convention for the Protection of Human Rights and Fundamental Freedoms states that "[e]veryone has the right to respect for his private and family life, his home and his correspondence" [24]. In 1981, the Council of Europe specifically addressed automated information collection and processing in its Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Data Convention). The Data Convention states that countries shall "respect … rights and fundamental freedoms, and in particular [the] right to privacy" for all individuals, regardless of nationality [25].

The 1949 UN Universal Declaration of Human Rights, Article 12, states that "No one shall be subjected to arbitrary interference with his privacy, family, home or

correspondence…" [26]. In 1980, the OECD issued Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (OECD Privacy Guidelines) [27]. These guidelines contain, in summary, the following principles: limitation of data collected, maintenance of data quality, specification of the collection purpose, limitation of data use to that specified purpose, adequate security, transparency, individual access to and control of data collected, and accountability. In 1998, the OECD reviewed the guidelines in light of the enormous changes in electronic communication, and confirmed the application of these basic principles to that environment. The OECD continues to work toward international standards in data privacy, and in recent work expanded its policy to the cross-border flow of information [28].

The current EU treatment of individual privacy builds upon these international and supranational foundational principles and policy documents, and is found primarily in two directives: the 1995 Directive on protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Directive) [29], and the 2002 Directive concerning the processing of personal data and the protection of privacy in the electronic communications sector [30]. The Data Directive created the Article 29 Working Party, or the Article 29 Board, as an independent advisory board on data protection. In 2007, the Article 29 Board issued the Working Document on the Processing of Personal Data Relating to Health in EHRs. The working document from Article Board provides an interpretation of the application of privacy principles to EHRs, and recommends adoption of 11 specific legal protections to protect individual health privacy. The report states unequivocally that "[a]ny processing of personal data in EHR systems has to fully comply with the rules for the protection of personal data." Hence, any data controller collecting individually identifiable health information must: limit data use to the purpose for which it was collected (purpose principle), ensure data quality (relevancy and accuracy), limit data retention (and not further process the data), provide individuals with data collection information and access to the information collected (with rights of correction), and provide appropriate data security measures (see Refs. [30] and [31]).

In general, the US legal framework for healthcare privacy is a combination of constitutional, statutory, and regulatory law at federal and state levels. While the EU and US privacy laws may seem to go about it in different ways, they share a common goal of protecting patients' privacy rights. Contrary to documented comparison that one is better than the other, it is not easy to come up with a definite answer, but each could learn a lot from the other. The regulatory framework for health care and privacy of other countries is similar to that in either the United States or the European Union (EU); therefore, in discussion in this chapter, the main focus will be on the United States and the EU [32].

The Privacy Act of 1974 [33] applies to federal agencies that hold individuals' personal information within any federal government records. The Health Insurance Portability and Accountability Act (HIPAA) is the primary law that establishes the US legal framework for health information privacy. Thus, the HIPAA privacy rule establishes standards for the privacy of individually identifiable health information and it applies to covered entities, defined as healthcare plans, healthcare providers, and clearing houses [34]. In addition to the privacy rule, the HIPAA security rule

requires covered entities to safeguard protected health information (PHI) through using administrative, technical, and physical measures. In 2009, HITECH was ratified to strengthen HIPAA's privacy and security guidelines [35] by imposing new privacy obligations on covered entities, expanding and clarifying business associates requirements; it also added provisions related to EHR, health information exchange and PHRs [36]. HITECH increased enforcement and monetary civil penalties.

In reality, many healthcare providers involve cloud providers or outside contractors to perform nonhealth functions such as record storage, processing, computer systems work, or billing. Those secondary entities can receive personal health information in the performance of their duties, and are addressed in the privacy rule under the category of "business associates." A "business associate" is "a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information" [34]. In order to disclose PHI to business associates, a covered entity must have assurances that the use of the information will be limited to that for which it was transferred, that the entity has sufficient security to protect the information, and that it will cooperate with the covered entity to protect the information as required under the privacy rule. Those assurances must be included in the agreement between the covered entity and the business associate [34].

Before HITECH, there was some confusion regarding whether an entity that processed information as a conduit, but was not using that information for other purposes, would be subject to the same privacy requirements as a covered entity. HITECH clarified and extended the regulation of business associates by providing that they are subject to the same privacy regulations applied to covered entities [35]. In summary, HITECH sought to put business associates under the same umbrella as covered entities in the protection of privacy and security of PHI.

The consent must be given by a positive communication of consent, an opt-in procedure rather than an insufficient opt-out procedure. In addition, the consent must be specific, voluntary, informed, and not coerced in any way. The opt-in consent might be used as a privacy-enhancing approach when collecting and processing sensitive data such as mental health treatment.

For users to accept and trust the EHR systems, both technical and legislation measures must be used. However, the legislative differences between countries erodes users' trust in EHR systems. For example, the US and EU assumptions for implementing privacy within an EHR system are worlds apart. One of the reasons is the differing foundations of individual privacy; whereas the United States established the right of an individual to health information privacy through a specific statute, in the EU this basic privacy right already existed under a human right to privacy, even before the adoption of the directives.

19.3.2 EHRs' Privacy International Collaborations in Standards and Protocols

The ISO/TS 13606-4:2009 seeks to address those requirements uniquely pertaining to EHR communications and to represent and communicate EHR-specific information that will inform an access decision. It also refers to general security requirements that

apply to EHR communications and points at technical solutions and standards that specify details on services meeting these security needs [15].

The ISO 27799 is a standard that has been specifically tailored to health care and that defines guidelines to support the interpretation and implementation in health informatics of ISO/IEC 27002. This standard addresses the information security management needs of the health sector. Implementing this guidance allows healthcare organizations to reduce the number and severity of their security incidents and to ensure a minimum level of confidentiality, integrity, and availability of personal health information. This standard provides clear, concise and healthcare-specific guidance on the selection and implementation of security controls for the protection of health information, and is adaptable to the wide range of sizes, locations, and service delivery models found in health care [14].

Other standards and regulations also appeared: the Role Based Access Control Standard of the "American National Standard for Information Technology" [37], The National Institute of Standards and Technology (NIST) RBAC reference model [38], the ISO TS 22600 "Health informatics—Privilege management and access control" [39], the ASTM E1986–98 "Standard Guide for Information Access Privileges to Health Information" [39], the ISO DTS 21298 "Health Informatics: roles of persons" [39], the "Recommendations for the Interpretation and Application of the Personal Information Protection and Electronic Documents Act" [40], the Recommendation R (75) in Europe "On the Protection of Medical Data" [41], the ENV 13729:2000 (Health informatics. Secure user identification. Strong authentication microprocessor cards), and the Privacy Code in New Zealand [41].

The ISO/IEC 27018 "Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors" builds on existing ISO standards such as ISO 27001 (the existing best practice for information security management) and is aimed at increasing confidence in data security and cloud computing. Adoption of the standard would complement one of the exceptions to the prohibition on transferring personal data outside of the European Economic Area (EEA) such as model contracts, binding corporate rules and safe harbor [42].

The ISO/IEC 27018 provides best practices for public cloud service providers (CSPs) and establishes guidelines for implementing measures to protect personal data. CSPs that adopt the standard agree to adhere to specific guidelines which include the following:

1. *Control*: only processing personal data in accordance with customers' instructions;
2. *Consent*: only processing personal data for marketing/advertising purposes with the customers' express consent (consent cannot be made a condition for receiving the cloud services);
3. *Communication*: notifying customers in the case of a breach and keeping clear records about the incident;
4. *Transparency*: disclosing to the customer the identity of sub-processors and any possible locations where personal data may be processed;
5. *Independent audit*: obtaining regular reviews of the CSP's compliance through a third-party independent audit.

CSPs that can market their services in accordance with the standard will provide greater consumer reassurance as to the quality of their services and address the downsides of cloud computing and the concerns of the cloud clients, mainly the lack of trust and transparency.

19.3.3 The Future of EHR Laws and Regulations

Privacy laws, regulations, standards, and policies will play a significant role in the provision of modern health care. They will shape and dictate the type of technologies used and how data (EHR) is stored, processed, and shared. This will mean more resources being used to come up with privacy laws, regulations, standards, and policies that will foster and not hinder innovation in healthcare provision.

There is a heated debate going on now on how much control should the patients be given in terms of managing their EHR by utilizing e-Health banks such as Microsoft HealthVault [43]. Systems that store, communicate, and process PHR on behalf of patients such as Microsoft HealthVault and Indivo [44] are not HIPAA covered entities. The PHR stored in these systems are fully owned by these companies. The patient and individual healthcare providers own portions of the medical records. The PHR in these systems are kept in proprietary format that are hard to integrate, audit, and exchange. In PHR privacy concerns such as information disclosure is going to be bigger than originally thought and maintaining accountability and auditing is not going to be easy. The challenge to legislators is to come up with laws that will restore the ownership of PHR to patients.

As EHR is digitized, transmitted, and mined for effective care provision, new forms of threats to privacy are becoming evident. In view of these emerging threats and the overarching goal of providing cost-effective healthcare services to all citizens, several important regulations have been enacted, including the privacy and security rules under HIPAA. The technology component involved in managing disclosure of EHR to third parties has led to stipulations of privacy compliance and provisions of security safeguards under HIPAA. The privacy rule of HIPAA addresses the use and disclosure of EHR by healthcare professionals, medical providers, and clearing houses also referred to as covered entities. By virtue of their contact with patients, covered entities are the primary agents of capturing EHR for various purposes including treatment, payment, managing healthcare operations, medical research, and subcontracting. New privacy laws, standards, regulations, and policies will be expected to improve the privacy protection offered to EHR by creating incentives and responsibilities to consumers and third-party users of healthcare data. This can be achieved by anonymizing EHR for purposes of sharing, establishing health information technology and privacy systems, bringing equity to healthcare provision, and increasing private enterprise participation in patient privacy.

As Web technology advances and connectivity becomes ubiquitous the quality, continuity, and convenience of offering and receiving health care will significantly improve. The cost of offering health care will be greatly reduced. Most of this will be achieved by collecting, storing, and processing data and metadata. However, provision of health care using Web technology will lead to collection of health-related information obtained by observing and inferring on Web queries. This is achieved by

third parties who are involved in tracking and sharing patient health-related information for profit. Prior research has demonstrated that while patients are uncomfortable with this type of tracking, it is performed in a number of highly sophisticated ways, and it is increasingly widespread. For example, most new technologies used in health care such as health mobile apps and medical devices are usually connected to the Internet, making them vulnerable to tracking and inferring attacks.

Another example of why privacy laws, regulations, standards, and policies might be an important part of the solution is the advertising companies that have developed massive data collection infrastructure that is designed to avoid detection, as well as ignore, counteract, or evade user attempts at limiting its collection. This means that a single company has the ability to record the Web activity of a huge number of individuals seeking sensitive health-related information without their knowledge or consent. As easy as it is to collect data, it is also easy for this information to be leaked and compromised. Other attacks involve the way poor data processing models may lead to wrong conclusions about the data affecting the privacy of users.

A lot of research in recent years has focused on developing technological solutions for ensuring privacy of patients when their information is stored, processed, and shared. It is equally important to recognize the contribution of technology adoption on continuity of care and care quality. Additionally, the enactment of new privacy laws, regulations, standards, and policies will be expected to improve patients' privacy. The emergence of Web-based healthcare applications while providing seamless portability of patient information may end up having detrimental impact on patients' privacy. This will definitely lead to proliferation of EHR identity theft and healthcare fraud.

19.4 PRIVACY OF EHRs IN e-HEALTHCARE SYSTEMS

In order to ensure the use and wide sharing of EHRs, mechanisms and tools for achieving privacy are required. This section discusses the mechanisms and tools for achieving EHR privacy showing their strengths and weaknesses.

19.4.1 Current EHR Privacy Solutions

Protection of the patients' privacy can be achieved using two different techniques, anonymization and encryption, which unfortunately both suffer from major drawbacks: While anonymization—the removal of the identifier from the medical data—cannot be reversed and therefore prevents primary use of the records by healthcare providers who obviously need to know the corresponding patient (patients cannot benefit from the results gained in clinical studies because they cannot be informed about new findings), encryption of the medical records prevents them from being used for clinical research (secondary use) without the explicit permission of the patient, who has to decrypt the data and, in doing so, reveals her/his identity. It must also be noted that encryption can be a very time-consuming operation [45]. A method that resolves some of these issues is

pseudonymization, where identification data are transformed and then replaced by a specifier that cannot be associated with the identification data without knowing a certain secret. Pseudonymization allows the data to be associated with a patient only under specified and controlled circumstances [45].

It is clear that the health care will increasingly be digitized in the future. This is facilitated among other things by the use of medical devices that are getting smarter and interconnected ensuring seamless communication within the e-Healthcare environment. For example, there are insulin pumps that send the data regarding usage to a computer via wireless connection. There are pacemakers that operate similarly. This leads to the question as to whether medical wireless communications can be compromised. Multiple news sources have carried a story of a researcher who discovered he could hack into wireless insulin pumps and alter the dosage, even to fatal levels. These devices are becoming increasingly complex and they collect, store, and transmit data in ways that are difficult to understand and track.

There have been several definitions for medical devices; in this chapter, we adopt two similar definitions by the World Health Organisation (WHO) [46] and the Food and Drug Administration (FDA) [47]. According to the FDA, a medical device is "an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article" intended for use in the diagnosis, prevention, monitoring, and treatment of disease or other conditions [47]. The WHO defines medical devices as a vast range of equipment, from simple tongue depressors to hemodialysis machines. Like medicines and other health technologies, they are essential for patient care at the bedside, at the rural health clinic or at the large, specialized hospital [46].

Medical devices can do some processing, receive inputs from outside the device (sensors), output values to the outer world (actuators), and communicate with other devices. Medical device security is the idea of engineering these devices so they continue to function correctly even if under malicious attack such as internal hardware and software aspects as well as intentional and unintentional external threats [48].

The connectivity, interoperability, and integration afforded to medical devices make their securing difficult. For example, major reasons for device failure include malfunction and software-related issues. Using communication protocol such as wireless further increases the attack surface. Securing medical devices means preventing malicious attacks taking control of the e-Healthcare infrastructure and ensuring the well-being of any person who needs to use medical devices.

Most of the software and hardware used in the hospital information systems infrastructure are the same as the ones used in nonmedical IT systems in hospitals and elsewhere. Therefore, the vulnerabilities and flaws are the same. And since there is a fine line distinguishing medical from nonmedical IT, that means nonmedical systems can pose threats to medical operations. For example, vulnerability in the Transmission Control Protocol/Internet Protocol (TCP/IP) stack can lead to distributed denial of service (DDoS) causing postponement of surgeries and patient treatment.

The use of medical devices involves using wireless communications protocols for sending and receiving data and instructions between the medical devices and the control unit. This makes the communication prone to the man-in-the-middle attack. Unless deliberate security measures are taken to protect the data on transition, we

have to assume it can be sniffed or eavesdropped by a malicious attacker. As sensor-rich medical devices become more ubiquitous, sensory malware has the potential to breach privacy of individuals at mass scale.

The availability and use of medical and mobile devices has led to increased use of self-caring services. Self-caring services are becoming more and more important in our daily life, especially as the number of old people globally increases. Self-caring helps reduce pressure in healthcare systems of many countries. Big data such as massive historical medical records makes it possible for users to have self-caring services, such as to get diagnosis by themselves with similar patients' records. Developing such a self-caring service gives rises to challenges including highly concurrent and scalable medical record retrieval, data analysis, as well as privacy protection [49].

A lot of medical treatments rely on software. The use of wired and wireless networking technologies in healthcare improves and/or adds functionalities in provision of health service and has also provided a space for innovation. However, as health care embraces the use of software and networking technologies, the unintended consequence is formidable security and privacy challenges. The process of getting medical devices approved is long and extremely complicated. It is extremely hard for a single person or a small company without deep pockets to undergo the lengthy medical devices and software approval process. The use of software in medical devices breeds overconfidence and may lead to a level of automation that may end up endangering patient health.

Wearable mobile devices will blanket the world. By 2025, there will be a massive Internet of everyone and everything linking every nation, community, company, and person to all of the world's knowledge. This will accelerate real-time access to education, health care, jobs, entertainment, and commerce. Artificial intelligence (AI) becomes both as smart as and smarter than humans. AI will be embedded in autos, robots, homes, and hospitals will create the AI economy. Humans and robots merge, digitally and physically, to treat patients who may be around the world. Robo-surgeons will operate remotely on patients. RoboDocs will deliver babies and treat you over the cell phone. Predictive medicine transforms health care. Early diagnosis of disease with medical devices that sniff our breath, and free DNA sequencing that predicts our future health will be common. Personalized genetic medicine will prevent disease, saving lives and billions in lost productivity [13].

The use of Cloud computing technology in health care is gaining popularity by facilitating an interorganizational medical data sharing environment (see Refs. [50], [51], and [52]). On the other hand, this paradigm also involves many security and privacy risks that lead to concerns among patients and medical workers (see Refs. [53], [54], and [55]) who are being particularly afraid of losing the control over sensitive medical records while storing them on not fully trusted third-party servers (see Refs. [56], [57], and [58]). The use of cloud computing in healthcare is becoming more important as facilitating interorganizational medical data sharing such as Microsoft HealthVault, WebMD, Georgia Tech's MedVault or Harvard's Indivo.

There are numerous examples of healthcare deployment in the cloud. In this chapter, we highlight the following projects: Loehr et al. [2] present a secure e-health

infrastructure based on Trusted Virtual Domains. Abbadi et al. [59] establish trustworthy middleware services with the goals of security, privacy, and resilience. Li et al. [60] elaborate on unlinkability between patients and medical records. References [54–57], and [61–63] leverage cryptographic access-control schemes to EHRs, while Ref. [60] enables a keyword search over encrypted documents.

Besides its benefits, the new promising cloud computing paradigm comes with many security and privacy risks that could impede its wide acceptance among patients and medical workers. The major concern is about losing control over sensitive medical records while storing them on a third-party server outside of the trust boundary of an organization. A primary concern for all stakeholders involved is the secure and privacy-preserving storage and handling of medical records (MRs). This involves the following major aspects.

Sensors can also be adopted for enabling different healthcare-related scenarios, for example, to automatically gather patient's data [64], to monitor hemoglobin concentration changes in the brain and tissues [65], to support the elderly in their self-activities and patients needing a physiological control [66], as well as to support patients suffering from depression [67].

19.4.2 Research and Practice into EHR Privacy

Privacy of patients and other users of e-Healthcare Systems is recognized as a fundamental requirement, but also one of the formidable challenges affecting the trust and acceptability of these systems. There are many proposals that have been put forward in the literature, most of which approach patient privacy as either an access control or an authentication problem. Currently, privacy has been looked at as a communication problem, since future e-Healthcare Systems will be highly distributed and require interoperability of many subsystems [68].

19.4.3 Trust Establishment Between the EHR, e-Healthcare Systems, Patients and Other Stakeholders

Tremendous amount of data are generated by the e-Healthcare Systems, which have special value when being shared and collaboratively used among different parties involved in the healthcare area (e.g., healthcare centers, laboratories, pharmacies, patients, health insurance, quality assurance in healthcare service delivery, researchers, national, and regional health authorities).

19.4.4 Privacy by Design in e-Healthcare Systems

Privacy is becoming a formidable challenge each passing day and it exists mainly in our encounter with technology. Major service providers like Facebook and Google and other similar, for example, have actively harvested data over the last number of years and they keep changing their policies to make use of the harvested data to deliver new personalized services to the customers. There are many solutions that have been implemented over the years to reduce breach of privacy. However, due to

increasing connectivity between new services and the inter-dependency between service providers is making the privacy management to be a challenging task for an innocent user. The same user who is a patient is faced with similar privacy management concerns when using e-Healthcare ICT and medical devices. These privacy challenges must be solved where possible or mitigated to make the e-Healthcare Systems safer and secure platforms to transact and communicate e-health information in the future.

In order to understand the difficult task of ensuring privacy in systems, the task of providing location privacy is used. According to Shokri [69], there are three entities that play a role in preserving location privacy: users, applications, and privacy tools. Each entity controls the amount of shared information and thus affects user privacy. Privacy policies influence the way applications can share information with different entities, and they are applied to the applications based on the user's decisions. Various privacy tools also use sophisticated algorithms to guarantee user's privacy. There are three main questions to be answered when we consider location privacy: What needs to be preserved? How can it be preserved? These two questions are the standard ones that all current research has been focused on. But the one question that is going to be the main concern now and in the future is "where to enforce privacy?"; there are three places where privacy can be enforced: the local network, the device, or the backend system. Most of the experimental reports in the literature suggest that the backend system is an optimal place to enforce privacy from performance and quality-of-output point.

Specifically, we know that medical devices do not have a lot of flexibility, bandwidth, and computation power. Besides all these, the devices do not have visibility of all other users in the network, and the backend system can perform computations faster and provide the results spontaneously to the device. The advantage of having the solution at the backend system is that it will have information about other people as well. The only catch with this solution is that trust with the backend system is needed. Latency with this solution might be higher than the device-based solution.

The current ongoing changes in digitization of the e-Healthcare Systems raises troubling issues regarding safety, privacy, and ethics. This has led to massive collection of data and ICT, leading to unknown challenges in medical information [70]. These changes in the way information is stored and retrieved have sensitized the public to issues of privacy such as healthcare workers sharing health records of famous people, identity frauds, or unintended disclosure of medical information through a stolen mobile device. There is understandably growing concern that these breaches of confidentiality may lead to an erosion of confidence in the e-Healthcare Systems in the eyes of the public, which may cause changes in the way individuals interact with the system. Such changes may lessen the actual quality of care provided, as patients can choose to self-medicate, see another doctor, give incomplete/inaccurate information or avoid seeking treatment to avoid disclosure.

For example, Powell et al. [71] surveyed British citizens concerning the introduction of the NHS national electronic health record system and inquired whether there were items and details that they preferred not to be shared with the national record

system. They found that the items participants tended to wish to keep to themselves were mostly related to pregnancy, contraception, sexual health, and mental health [71]: that is, information considered sensitive, private, and potentially stigmatizing. As has been observed in several recent studies [72], simply getting people to disclose medical information that would be digitally stored and processed depends on a complex array of factors and circumstances, including culture, their emotional response, and several mechanical conditions and contextual factors.

Healthcare professionals have privacy and confidentiality concerns as well, which may hinder the acceptance and adoption of EHR [73]. In light of these troubling findings and the growing awareness of potential violations of privacy, the health system has begun to employ several methods to protect the privacy of its clients. Several e-Healthcare Systems have adopted security standards (i.e., the 1996 HIPAA and EU Health Directive) [14]. Another method that is being used in some cases is pseudo-anonymity [45]. The concept of pseudo-anonymity allows third parties (i.e., clinics, hospital point of care, and researchers) to access medical data without disclosing the patient's identity (i.e., using identifier instead of a full name).

Other measures to protect the privacy of patients are in the form of access control. Several e-Healthcare Systems require authorization and authentication procedures such as electronic signatures, anonymous digital credentials, smart cards, or limited access policies [14, 74].

19.4.5 Design of Custom Tools and Protocols for Monitoring and Managing Privacy of EHR and e-Healthcare Systems

The use of mobile devices and especially mobile apps is a welcome addition in allowing patients collect healthcare related data and conveniently access e-Healthcare Systems. Mobile apps are usually installed on the client side; therefore, they are under the user's control. Mobile apps can be easily compromised and can cause severe health hazard to the patient. For example, malicious users or competitors could decompile the application and analyze the result. This could reveal valuable data as proprietary algorithms or intellectual property or allow the attackers to use that information to modify the code, repackage, and redistribute it to create a "trojanized" clone of the app in a rapid fashion. Moreover, if the app needs to run on untrusted devices, any malware could interact with the app at runtime level to steal data (credentials, credit card number) or bypass security logic (local authentication, geo-restrictions, custom cryptography, etc.). As can be seen mobile Apps could be attacked at various layers and with very different goals in mind, creating a very complex problem for e-Healthcare Systems protection.

Some of the techniques that can help mitigate security and privacy threats in mobile devices include the implementation of the following techniques: Code and Flow Obfuscation, String and Class Encryption, Debug code stripping, Method Call Hiding (Reflection), Resource Encryption, Debug Detection, Root/Jailbreak Detection, Runtime Injection Detection (Swizzle/Hook Detection), Tamper Detection, Certificate Pinning, and Watermarking [75].

The distinction between empowerment and protection could also inform discussions about technical privacy defense tools. For example, technology could help foster meaningful transparency regarding data processing and profiling, and user-friendly mechanisms needed to give, withhold, or retract consent. However, in some circumstances, people might benefit more from being confronted with transparency and choices. Examples of more protective technical approaches include services that automatically secure personal information, metadata, or communications regardless of the user's initiative [76].

There are a number of ongoing efforts to restore and build user's trust in mission critical systems such as e-Healthcare Systems. One of the promising frameworks is the Fast IDentity Online (FIDO) [77]. The FIDO framework aims to address the lack of interoperability among strong authentication devices as well as the problems users face with creating and remembering multiple usernames and passwords. It plans to change the nature of authentication by developing specifications that define an open, scalable, interoperable set of mechanisms that supplant reliance on passwords to securely authenticate users of online services. This new standard for security devices and browser plugins will allow any website or cloud application to interface with a broad variety of existing and future FIDO-enabled devices that the user has for online security [77].

19.5 DISCUSSION AND CONCLUSION

There is an ongoing debate among security and privacy experts, consumer advocates, and the healthcare professionals about the extent of access and control patients should have over their EHR. Some contend that patient policies that require too much patient control "could hamper a patient's health in a medical emergency," while on the other side, it is "said that not enough control could put their lives at risk in other ways" [78]. Patient Privacy Rights, an advocacy organization dedicated to ensuring that Americans control access to their health records, argues that "[a] lack of safeguards … poses risks to a person's well-being, livelihood and financial stability," and because of this "patients should have total control over their personal health records … to ensure that information that a patient wishes to be kept private is kept that way." On the other hand, the Center for Democracy and Technology's Health Privacy Project argues, "patients intuitively want control of their data, but requiring consent for every exchange of health information is sometimes not the best approach for ensuring privacy" [78]. However, all agree that without patient's trust in the security and privacy of their EHR, most e-Healthcare Systems will fail.

Global epidemics and emergencies such as the SARS crisis of 2003 and the Avian Flu threat demonstrate both the reality of a new disease spreading throughout a "highly mobile, closely interconnected and interdependent world" as well as the limitations of national governments in responding to such challenges independently [79]. These global epidemics realities make the challenges having a cloud based e-Healthcare Systems worthy solving.

19.6 CONTRIBUTIONS AND FUTURE RESEARCH

In this chapter, we recommend that more research should be undertaken in order to accommodate new technologies as they find their way into the e-Healthcare landscape so that adopted regulations, standards, and technical measures could lead to the deployment of secure EHR systems. A huge challenge will be to understand and characterize the level of complexity of such a world in which many billions of devices are interacting with one another often in unplanned ways.

In order to ensure continuity of care and development of ICT in the provision of health care, EHR must be shared. Sharing allows EHR to be used for treatment, policy, and research. However, this level of sharing possesses security threats as defects in some of e-Healthcare Systems could cause the disclosure of information to unauthorized persons or companies. It is important that health data are protected against manipulations, unauthorized accesses, and abuses, which includes taking into account privacy, trustworthiness, authentication, responsibility, and availability issues. The challenges of ensuring the privacy of EHR are magnified due to the sharing dimension of EHR.

The secondary use of EHR has a vital role in improving and advancing medical knowledge. While EHR lends itself nicely to sharing, it is essential to respect patient's wishes regarding secondary usage, and ensure the privacy of the patient. Consent, together with depersonalization and its related concepts of anonymization, pseudonymization, and data minimization are key methods used to provide this protection [3].

Cloud computing is a new computing paradigm for providing large-scale computational resources everywhere. Its usefulness has come to scrutiny in recent times due to constant stream of security and privacy breaches. Storing and sharing of medical data in the cloud environment, where computing resources including storage is provided by a third-party service provider, raise serious concern of individual privacy for the adoption of cloud computing technologies [80]. In this chapter, we argue that the future of the cloud requires fundamental changes in the way large-scale distributed systems are designed to account for data confidentiality and privacy. For example, one of the solutions proposed in the cloud space is the use of encryption, which comes at a higher cost of key management. There are several innovative proposals in how to handle the problem of key management.

It may come as a surprise that encryption does not solve the confidentiality problem. It just shifts the problem from preserving the data's confidentiality to preserving the encryption keys' confidentiality. Because encryption keys cannot be trusted by servers (cloud servers), key distribution and revocation are nontrivial problems, which in turn generate high management costs. Some systems try to solve the problem of key management by employing secret sharing to store the keys, effectively running a secret sharing-based system to bootstrap the cryptography-based system [81].

There is light at the end of the tunnel with the near realization of being able to process encrypted information. Processing encrypted data is orders of magnitude slower that processing plaintext, although some of the recent results have shown that it might eventually become practical [82].

The underlying ICT systems, medical, and mobile devices are using consumer grade hardware and software. This means that they are vulnerable to all security and privacy threats as other systems plus the protection and sharing while ensuring privacy of EHR. The other major problem is the fragmentation of the EHR depending on where the patient has accessed care, with no rapid system of cross-referencing or integration of such records being available. The current advances and use of technologies in health care is bound to make the EHR fragmentation problem worse.

At the moment, most projects in health care are localized, fragmented, distributed, and small. However, tomorrow's projects that will be a confluence of EHR, ICT, medical and mobile devices, connectivity, and regulations will be larger, smarter, and global in scale. These projects will be integrated and interoperable at a level no one could have imagined today. When we reach such a stage, the government leadership will be inevitable, as the projects will be larger for single or several vendors to handle. The government leadership will be required in the area of designing open standards and ensuring compliance in its usage.

We are now becoming aware that all technologies we use are not the most important bit of the puzzle although they have advanced significantly. It's what the society does with them, and right now it's institutional change that's the sticking point. What you really want to look at is new ways of organizing ourselves. For example, the Open Source Drug Discovery (OSDD) which is a consortium with global partnership with a vision to provide affordable health care to the developing world by providing a global platform where the best minds can collaborate and collectively endeavor to solve the complex problems associated with discovering novel therapies for neglected tropical diseases such as tuberculosis, malaria, and leishmaniasis [83]. A free patient network like PatientlikeMe is one where people can connect with others who have the same disease or condition and track and share their own experiences. When people share information on PatientslikeMe website, they generate data about the real-world nature of disease that helps researchers, pharmaceutical companies, regulators, providers, and nonprofits develop more effective products, services, and care [84].

In order to ensure the privacy of EHR, laws play a very important role. For example, the European Data Protection guidelines that has recently undergone revisions to include the privacy of individual's data and personally identifiable information (PII). Other notable changes include explicit consent from the user when data are being shared with other third-party service providers. More transparency about the way in which the data are handled is another important change to European Data Privacy Directive. The reform also includes the mandate for complete accountability and responsibility of the service provider when personal data are being processed [29].

As ICT becomes an integral part of the e-Healthcare System, mobile devices will be used increasingly. While mobile devices to be used in this environment are not subjected to level of scrutiny that medical devices are, they need to prove their worth either way. Mobile devices will bring convenience and usability to the user, but also all the vulnerabilities and attacks associated with it. In other words as e-Health becomes more heterogeneous, its attack surface will increase considerably, affecting the privacy of data and its users.

The issue of e-Healthcare is constantly accompanied by the need to develop appropriate assessment methods, security and privacy mechanisms, and protocols. Such assessment tools may also prove useful in the implementation and dissemination process of future e-Healthcare projects.

Uploading and sharing information in a digital process raise legal questions concerning the protection of information, shortcuts in care, and potential increases in medical malpractice lawsuits. It is clear that patients do not want their EHR shared and doctors are more reluctant to use the system because of legal concerns and doctor–patient confidentiality. These issues are crucial in the design and adoption of the system, and the EHR must protect both patient's medical records and doctor integrity.

There are a lot of mobile devices and apps that are providing health-related services from monitoring blood pressure to counting the calories consumed in an exercise. These devices and apps are being designed and developed with developers who have no to little medical training and without the rigor required for devices or apps that collect sensitive EHR. It is not easy to stop or fully regulate this burgeoning industry of producing health-related devices and apps. It is extremely worrying that users rely and accept results produced by them, and undue pressure is exerted on healthcare professionals to accept the results. It is just a matter of time before data collected by mobile devices and apps form a significant segment of EHR.

The on-demand revolution will become the on-demand world, where biological software upgrades, personalized medicine, and AI assistants will increasingly transform health care and well-being. Additionally, increased automation will continue to make our day-to-day lives infinitely richer. Self-driving cars will be ubiquitous; transportation itself will be automatic, clean, and cheap. We will move into a world in which access trumps ownership and the world is at our fingertips [13].

We have reached a stage where in order to maintain and enjoy the level of development in medical area our sensitive private data will have to be used as the crude oil of this revolution. Current research, design, and development of new products should be aimed at delivering personalized medical services. This should be achieved by selectively collecting and processing our personal data securely to afford privacy protection.

REFERENCES

[1] OECD. Why is the health spending in the United States so high? 2012. Available at: http://www.oecd.org/unitedstates/49084355.pdf (accessed on May 20, 2015).

[2] Loehr H., Sadeghi A.R., Winandy M. Securing the E-health cloud. Proceedings of the First ACM International Health Informatics Symposium; November 11–12, 2010, Arlington, VA, USA.

[3] Elger B.S., Iavindrasana J., Iacono L.L., Müller H., Roduit N., Summers P., et al. Strategies for health data exchange for secondary, cross-institutional clinical research. Computer Methods and Programs in Biomedicine. 2010; 99(3): 230–251.

[4] Zheng Y.L., Ding X.R., Yan Poon C.C., Lai Lo B.P., Zhang H., Zhou X.L., et al. Unobtrusive sensing and wearable devices for health informatics. IEEE Transactions on Biomedical Engineering. 2014; 61(5): 1538–1554.

[5] Murdoch T.B., Detsky A.S. The inevitable application of big data to health care. Journal of the American Medical Association. 2013; 309(13): 1351–1352.

[6] AbuKhousa E., Mohamed N., Al-Jaroodi J. e-Health cloud: Opportunities and challenges. Future Internet. 2012; 4: 621–645.

[7] Jha A.K., Adler-Milstein J. Regional Health Information Organizations and Health Information Exchange. In: Blumenthal D., ed. Health Information Technology in the United States: Where We Stand. BMJ Publishing Group Limited, Harvard University, Cambridge, MA, 2008, p. 8.

[8] Donelan K., Miralles P.D. Consumers, EHRS and PHRs: Measures and Measurement. In: Blumenthal D., ed. Health Information Technology in the United States: Where We Stand, Robert Wood Johnson Foundation, Princeton, NJ, 2008, pp. 56–57.

[9] Bernat J.L. Ethical and quality pitfalls in electronic health records. Neurology. 2013; 80 (11): 1057–1061.

[10] Mangalmurti S.S., Murtagh L., Mello M.M. Medical malpractice liability in the age of electronic health records. New England Journal of Medicine. 2010; 363(21): 2060–2067.

[11] Kisilev P., Walach E., Barkan E., Ophir B., Alpert S., Hashoul S.Y. From medical image to automatic medical report generation. IBM Journal of Research and Development. 2015; 59(2/3): 2:1–2:7.

[12] WHO. The global burden of chronic. 2015. Available at: http://www.who.int/nutrition/topics/2_background/en/ (accessed on May 9, 2015).

[13] Howard J. 7 Top futurists make some pretty surprising predictions about what the next decade will bring. 2015. Available at: http://www.huffingtonpost.com/2015/05/12/futurists-next-10-years_n_7241210.html (accessed on May 9, 2015).

[14] Fernández-Alemán J.L., Señor I.C., Lozoya P.Á.O., Toval A. Security and privacy in electronic health records: A systematic literature review. Journal of Biomedical Informatics. 2013; 46(3): 541–562.

[15] ISO. ISO/EN 13606. 2012. Available at: http://www.iso.org/iso/home.htm/ (accessed on May 15, 2015).

[16] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002; 10(5): 557–570.

[17] Loukides G., Denny J.C., Malin B. The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association. 2010; 17: 322–327.

[18] Atreya R.V., Smith J.C., McCoy A.B., Malin B., Miller R.A. Reducing patient re-identification risk for laboratory results within research datasets. Journal of the American Medical Informatics Association. 2013; 20(1): 95–101.

[19] Sandhu R.S., Cope E.J., Feinstein H.L., Youman C.E.. Role-based access control models. IEEE Computer. 1996; 29(2): 38–47.

[20] Pinkas B. Cryptographic techniques for privacy-preserving data mining. ACM Special Interest Group on Knowledge Discovery Data Explorations Newsletter. 2002; 4(2): 12–19.

[21] Loukides G., Liagouris J., Gkoulalas-Divanis A., Terrovitis M.. Disassociation for electronic health record privacy. Journal of Biomedical Informatics—Special Issue on Informatics Methods in Medical Privacy. 2014; 50: 46–61.

[22] Meystre S.M., Escamez Ó.F., Jeffrey F.F., South B.R., Shen S., Samore M.H. Impact of de-identification on clinical text information content. Journal of Biomedical Informatics. 2014; 50: 142–150.

[23] Simpson C. Policy as a process: The pedagogical role of the EU in health care. North Carolina Journal of International Law and Commercial Regulation. 2007; 33: 293–335.

[24] Convention for the Protection of Human Rights and Fundamental Freedoms, art. 8, November 4, 1950, *213 UNTS 222*.

[25] EU. Convention for the protection of individuals with regard to automatic processing of personal data. 1981. Available at: http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm (accessed on May 14, 2015).

[26] Universal Declaration of Human Rights, G.A. Res. 217A, at 71, U.N. GAOR, 3d Sess., 1st plen. mtg., U.N. Doc. A/810 (December 12, 1948).

[27] OECD. Guidelines on the protection of privacy and transborder flows of personal data (Paris 1981). 1981. Available at: http://www.oecd.org/sti/ieconomy/oecdguidelineson theprotectionofprivacyandtransborderflowsofpersonaldata.htm (accessed on May 15, 2015).

[28] OECD. New revised guidelines on the protection of privacy and trans border flows of personal data (2013). 2013. Available from: http://www.oecd.org/sti/ieconomy/privacy.htm (accessed on May 14, 2015).

[29] Council Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data, *art. 30*.

[30] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). *Council Directive 2002/58/EC art. 5, 6, 9*.

[31] Article 29 Data Protection Working Party, Working Document on the Processing of Personal Data Relating to Health in Electronic Health Records 2 (Article 29 Working Party). *Working Paper No. 131, 2007*.

[32] Janine H., McMullen M.S., Chumney W.M., Baumer D.L.. Privacy and security in the implementation of health information technology (electronic health records): U.S. and EU compared. Journal of Science & Technology Law [Internet]. 2011; 17(1): 1–40.

[33] Privacy Act 1974. *Pub L No 93-579*, 88 Stat. 1896, enacted December 31, 1974, 5 U.S.C. A federal USA law.

[34] Health Insurance Portability and Accountability Act of 1996. *Pub L No 104-191*.

[35] Health Information Technology for Economic and Clinical Health (HITECH) Act. *Pub L No 111-5*.

[36] Manos D. HHS issues rule on EHR breach notification. 2009. Available at: http://www.healthcareitnews.com/news/hhs-issues-rule-ehr-breach-notification (accessed on May 14, 2015).

[37] Røstad L. An initial model and a discussion of access control in patient controlled health records. The Third International Conference on Availability, Reliability and Security ARES; March 4–7, 2008, Technical University of Catalonia, Barcelona, Spain, pp. 935–942.

[38] Ferraiolo D.F., Sandhu R., Gavrila S., Kuhn R.D., Chandramouli R. Proposed NIST standard for role-based access control. ACM Transactions on Information and System Security (TISSEC). 2001; 4(3): 224–274.

[39] van der Linden H., Kalra D., Hasman A., Talmon J. Inter-organizational future proof EHR systems. A review of the security and privacy related issues. International Journal of Medical Informatics. 2009; 78(3): 141–160.

[40] Narayan S., Gagné M., Safavi-Naini R. Privacy preserving EHR system using attribute-based infrastructure. Proceedings of the ACM Workshop on Cloud Computing Security; October 4–8, 2010, Chicago, IL, USA, pp. 47–52.

[41] Hembroff G.C., Muftic S. SAMSON: Secure access for medical smart cards over networks. Proceedings of the IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM); June 14–17, 2010, Montreal, Canada, pp. 1–6.

[42] ISO. ISO/IEC 27018:2014 Information technology—Security techniques—Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors. 2014. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=61498 (accessed on May 9, 2015).

[43] Microsoft. Microsoft HealthVault. 2015. Available at: https://www.healthvault.com/ie/en (accessed on May 20, 2015).

[44] Indivo. Indivo. 2015. Available at: http://indivohealth.org/ (accessed on May 20, 2015).

[45] Neubauer T., Heurix J.. A methodology for the pseudonymization of medical data. International Journal of Medical Informatics. 2011; 80(3): 190–204.

[46] WHO. Medical device regulations: Global overview and guiding principles. World Health Organization, Geneva, 2003.

[47] FDA. Is the product a medical device? 2012. Available at: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/ucm051512.htm (accessed on April 20, 2015).

[48] Sametinger J., Rozenblit J., Lysecky R., Ott P. Security challenges for medical devices. Communications of the ACM. 2015; 58(4): 74–82.

[49] Lin W., Dou W., Zhou Z., Liu C.. A cloud-based framework for home-diagnosis service over big medical data. Journal of Systems and Software. 2015; 102: 192–206.

[50] Fabian B., Ermakova T., Junghanns P.. Collaborative and secure sharing of healthcare data in multi-clouds. Information Systems. 2015; 48: 132–150.

[51] Shini S., Thomas T., Chithraranjan K. Cloud based medical image exchange-security challenges. Proceedings of International Conference on Modelling, Optimization and Computing; March 5–9, 2012, Hanoi, Vietnam, pp. 3454–3461.

[52] Guo L., Chen F., Chen L., Tang X.. The building of cloud computing environment for E-health. Proceedings of International Conference on E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference on Technologies; April 2010, Shenzhen, China.

[53] Li M., Yu S., Cao N., Lou W. Authorized private keyword search over encrypted personal health records in cloud computing. Proceedings of 31st International Conference on Distributor Computer Systems; June 20–24, 2011, Minneapolis, MN.

[54] Li M., Yu S., Zheng Y., Ren K., Lou W. Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. IEEE Transactions on Parallel and Distributed Systems. 2012; 24(1): 131–143.

[55] Chen T.S., Liu C.H., Chen T.L., Chen C.-S., Bau J.G., Lin T.C. Secure dynamic access control scheme of PHR in cloud computing. Journal of Medical Systems. 2012; 36(6), 4005–4020.

[56] Chen L., Hoang D.B. Novel data protection model in healthcare cloud. Proceedings of IEEE International Conference on High Performance Computing and Communications; September 13–15, 2011, Banff, Canada.

[57] Li M., Yu S., Ren K., Lou W. Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings. Proceedings of SecureComm 2010, LNICST; September 7–9, 2010, Singapore, pp. 89–106.

[58] Akinyele J.A., Lehmann C.U., Green M.D., Pagano M.W., Peterson Z.N.J., Rubin A.D. Self-Protecting Electronic Medical Records Using Attribute-Based Encryption; 2010. Cryptology ePrint Archive, Report 2010/565. Available at: http://eprint.iacr.org/ (accessed on February 22, 2016).

[59] Abbadi I.M., Deng M., Nalin M., Martin A., Petkovic M., Baroni I. Trustworthy middleware services in the cloud. Proceedings of Third International Workshop on Cloud Data Management; October 24–28, 2011, Glasgow, UK.

[60] Li Z.R., Chang E.C., Huang K.H., Lai F. A secure electronic medical record sharing mechanism in the cloud computing platform. Proceedings of IEEE 15th International Symposium on Consumer Electronics; June 14–17, 2011, Singapore.

[61] Basu S., Karp A., Li J., Pruyne J., Rolia J., Singhal S., et al. Fusion: Managing healthcare records at cloud scale. IEEE Computer, 2012; 45(11): 42–49.

[62] Chen Y.Y., Lu J.C., Jan J.K. A secure EHR system based on hybrid clouds. Journal of Medical Systems. 2012; 36(5): 3375–3384.

[63] Yu S., Wang C., Ren K., Lou W. Achieving secure, scalable, and fine-grained data access control in cloud computing. Proceedings of 29th Conference on Information Communications; March 2010, San Diego, CA, USA.

[64] Rolim C.O., Koch F.L., Westphall C.B., Werner J., Fracalossi A., Salvador G.S. A cloud computing solution for patient's data collection in health care institutions. Proceedings of the Second Telemedicine and Social Medicine; February 10–16, 2010, St. Maarten, Netherlands Antilles.

[65] Sharieh S., Franek F., Ferworn A. Using cloud computing for medical applications. Proceedings of 15th Communications and Networking Simulation Symposium; March 26–29, 2012, Orlando, FL, USA.

[66] Berndt R.D., Takenga M.C., Kuehn S., Preik P., Sommer G., Berndt S. SaaS—platform for mobile health application. Proceedings of Ninth International Multi-Conference on Systems Signals and Devices; Mar 20–23, 2012, Chemnitz University of Technology, Germany.

[67] Deng M., Nalin M., Petkovié M., Baroni I., Marco A. Towards trustworthy health platform cloud. Proceedings of IEEE Fourth International Conference on Cloud Computing; October 2011; 2012, Glasgow, UK.

[68] Dong N., Jonker H., Pang J.. Challenges in eHealth: From Enabling to Enforcing Privacy. Foundations of Health Informatics Engineering and Systems, Lecture Notes in Computer Science. 2012; 7151: 195–206.

[69] Shokri R. Quantifying and Protecting Location Privacy. École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; 2013.

[70] Malin B.A., El Emam K., O'Keefe C.M. Biomedical data privacy: Problems, perspectives, and recent advances. Journal of the American Medical Informatics Association. 2013; 20(1): 2–6.

[71] Powell J., Fitton R., Fitton C. Sharing electronic health records: The patient view. Informatics in Primary Care. 2006; 14(1): 55–57.

[72] Anderson C.L., Agarwal R. The digitization of healthcare: Boundary risks, emotion, and consumer willingness to disclose personal health information. Information Systems Research. 2011; 22(3): 469–490.

[73] Chao W.C., Hu H., Ung C.O., Cai Y. Benefits and challenges of electronic health record system on stakeholders: A qualitative study of outpatient physicians. Journal of Medical Systems. 2013; 37(4): 1–6.

[74] Blobel B. Authorisation and access control for electronic health record systems. International Journal of Medical Informatics. 2004; 73(3): 251–257.

[75] Antitamper mobile—Minded security's magik quadrant for mobile code protection. 2015. Available at: http://blog.mindedsecurity.com/2015/04/antitamper-mobile-minded-securitys.html (accessed on May 15, 2015).

[76] Borgesius F.Z.. Informed consent: We can do better to defend privacy. IEEE Security and Privacy. 2015; 13(2): 103–107.

[77] FIDO. Fast IDentity Online (FIDO). 2015. Available at: https://fidoalliance.org/ (accessed on May 15, 2015).

[78] Manos D. Privacy experts debate patient consent. 2009. Available at: http://www.healthcareitnews.com/news/privacy-experts-debate-patient-consent (accessed on May 14, 2015).

[79] WHO. WHO Issues Paper of 3 November 2003, MIP/2003/IP/1; 2003.

[80] Yang J.-J., Li J.-Q., Niu Y.. A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Generation Computer Systems 2015; 43–44: 74–86.

[81] Reiter M.K., Franklin M.K., Lacy J.B., Wright R.N.. The omega key management service. ACM Conference on Computer and Communications Security; March 14–15, 1996, New Delhi, India, pp. 38–47.

[82] Gentry C.. Fully homomorphic encryption using ideal lattices. Proceedings of the 41st Annual ACM Symposium on Theory of Computing; May 31, 2009, Bethesda, MD, USA, pp. 169–178.

[83] Open Source Drug Discovery. 2015. Available at: http://www.osdd.net/home (accessed on May 15, 2015).

[84] PatientsLikeMe. PatientsLikeMe. 2015. Available at: http://www.osdd.net/home (accessed on May 15, 2015).

20

ETHICAL, PRIVACY, AND INTELLECTUAL PROPERTY ISSUES IN NANOMEDICINE

Katie Atalla[2], Ayush Chaudhary[1], Mary Mehrnoosh Eshaghian-Wilner[2], Arushi Gupta[2], Raj Mehta[2], Adarsh Nayak[2], Andrew Prajogi[2], Kodiak Ravicz[2], Ben Shiroma[2], and Pujal Trivedi[2]*

[1] *Intel, Santa Clara, CA, USA*

[2] *EE-Systems, University of Southern California, Los Angeles, CA, USA*

20.1 INTRODUCTION

The field of medicine is on the verge of a revolution caused by the implementation of the first nanotechnology-based treatments and medicines. Over 200 companies worldwide are engaged in nanomedicine research and development (R&D), and the vast majority of these are startups or small- to medium-sized enterprises rather than large pharmaceuticals [1]. Furthermore, the US Food and Drug Administration (FDA) has already approved nine nanotechnology-based therapies, with applications ranging from early detection medical imaging to regenerative bone and tissue repair [2]. By taking advantage of the sub-micron-level material properties of select compounds, researchers are able to develop more precise, effective, and easily administered treatments [3]. Nanomedicine has revolutionary potential; it brings us closer to achieving many of the United Nations' (UN) Millennium Development Goals and holds great promise for advancing health care in developed and developing countries alike [4]. However, for every innovation nanomedicine promises, the unknowns within this Pandora's box threaten individuals, society, and even the environment as a whole.

*Authors are listed alphabetically by last name, as opposed to by order of contribution.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

The very untested nature of nanomedicine that makes it so revolutionary also makes it problematic and potentially dangerous. This new technology gives rise to several ethical, privacy, intellectual property (IP), and legal issues. For example, can a nanomedicine trial participant's informed consent be valid, when even the researchers conducting the study do not know the full extent or magnitude of the risks involved [5]? Are e-healthcare systems secure enough to make the personalized medical data from e-healthcare system integrated, nanotechnology-based medical devices worth the risk of unauthorized access or use (see Chapter 19)? Can the IP law system adequately protect nanomedicine without fostering an environment that encourages abuse, which ultimately stymies innovation? And are existing institutions even capable of regulating nanomedicine, a task that falls under the—often conflicting—jurisdictions of several different regulatory agencies [5]?

Ultimately, these issues will not be addressed in any comprehensive manner without cooperation between the public, researchers, pharmaceutical and biotechnology companies, and regulatory agencies. The ethical issues posed by nanomedicine are too important to be dismissed without significant discourse between researchers and the affected public. Privacy issues will only be adequately solved if they are addressed in a bottom-up fashion from the biotechnology companies and healthcare providers themselves, instead of a piecemeal top-down legislative solution from regulators. There is no more clear an example of legislation without industry input than the ailing IP system in which ineffective IP laws have been upheld despite objections [6]. If nanomedicine innovations are to be protected without preventing future progress, a dialogue between legislators and industry needs to be established. Similarly but even more importantly, if nanomedicine is to be regulated to the extent that the potential for harm the technology possesses demands, the public, industry, and legislatures all must work in tandem to create a regulatory body capable of handling the unique challenges posed by nanomedicine.

The nontechnical issues that nanomedicine poses can be divided into roughly four categories. While this list is not wholly dispositive and undoubtedly many readers' analysis will produce additional issues raised by this emerging technology, these four categories both exemplify and summarize the major ethical, privacy, IP, and other legal issues surrounding nanomedicine. Ethical issues are broken down by where they arise in a nanotechnology-based treatment's lifecycle from R&D, to distribution, to its final application and use. To clarify these issues, existing and developed medical ethics paradigms are used in this analysis. The privacy issues discussed in previous chapters are summarized here and further expanded upon. IP issues while technically a subset of legal issues are discussed separately given their importance in the initial development and evolution of a new technology. Subsequently other legal issues, particularly the convergence of jurisdiction of various regulatory bodies on nanomedicine is investigated, and previously discussed privacy laws governing patient data are recapped.

20.2 ETHICAL ISSUES

Ethical considerations should play a central role in the development and deployment of nanomedicine and guide scientists in their research of this fledgling technology. The Woodrow Wilson Center outlined the challenge posed by nanomedicine and all new technologies.

> Scientists have given and will continue to give us vast marvels, capable of producing technologies of great power. Each of these marvels, including nano[medicine], comes in a treasure chest of riches and a Pandora's box of evils. The challenge… is to use the treasure while keeping shut the lid on Pandora's box. It is a daunting challenge, but one that can be met. [7]

Ethical discourse is the tool that enables society to distinguish the promise of new technologies from dangers they pose. Unfortunately, this has not been the case with nanomedicine. As Peter Singer writes "as [nanomedicine] leaps ahead, the ethics lags behind" [8]. He further warns that nanotechnology and all of its dependent fields could be "derail[ed]… if the study of ethical, legal, and societal implications does not catch up with the speed of scientific development" [8]. However, it is imperative the discussion remains firmly rooted in the actual science of nanomedicine and resists the tendency to veer toward either extreme, presenting nanomedicine as either salvation or damnation [9]. An informed ethical discussion guiding the research and use of nanomedicine is necessary for the responsible utilization of this technology and to reach its maximum societal potential.

In this section, first, the question of what is nanoethics will be addressed. Second, what it means for an ethical issue to be truly "new" will be investigated. This framework will be applied to these issues of nanoethics. Third, parallels between these issues and existing issues of applied ethics will be drawn as well as broken down into five main conceptual categories by which the narratives they appeal to can be understood. Fourth, a nanotechnology-based treatment's lifecycle is divided into R&D, distribution, and final application stages. Last, these stages are used to categorize specific illustrative examples of ethical issues caused by nanomedicine.

20.2.1 What Is Nanoethics?

A common, albeit self-explanatory definition of nanoethics is that it is "the study of the ethical and social dimensions of nanotechnology" [10]. This definition is problematic or at the very least interesting for a few reasons. The first reason is that as previously discussed the definition of nanotechnology is ill-defined and relatively malleable, and thus any derivative definitions are equally if not more ambiguous. A second reason is that the field of nanoethics is inclusive of the social and even legal issues surrounding nanotechnology [10]. This inclusive definition is possibly due to the mingled nature of the impacts nanotechnology is having and will have on the world. However, while this implicit inclusion is often assumed in discussions of nanoethics, it must be noted that the topics of ethical, social, and legal impacts are sometimes referred to separately, but contemporaneously as well with groupings like social and ethical interactions with nanotechnology (SEIN); nanotechnology's ethical, legal, and social implications NELSI; or nano-ethical, environmental, economic, legal, and social issues (NE^3LS) [11, 12]. Despite these difficulties in defining nanoethics and while there is some debate among scholars as to whether nanoethics truly constitutes its own branch of applied ethics, nanoethics as it refers to the ethical issues encountered while engaging in the research, production, distribution, or use of

nanotechnology and by extension nanomedicine, is important due to the gravity of the issues raised and the magnitude of their consequences.

20.2.2 Types of Ethical Issues Associated with Nanomedicine

As this text does not have a true ethics focus but rather a technical one supplemented by an informed discussion of surrounding ethical issues, the ethical issues associated with nanomedicine are best organized here by where in the nanotechnology-based treatment's lifecycle they occur. This has the benefit of avoiding most of the field of ethics' byzantine and esoteric terms that add little of value to this discussion. Conversely, this stage-based approach has the benefit of an intuitive ordering as it follows a chronological progression. Furthermore, most of the issues that occur during the same stage share a common ethical theme, whether it is incomplete risk analysis stemming from untested nanomedicines in the R&D stage; inequities of access and need in the distribution stage; or normative questions of how nanomedicines should be used in the application and final use stage generally all come back to the treatment or enhancement debate [10].

The R&D stage is by all premarket research, laboratory testing of materials and human and animal testing that a nanotechnology-based treatment undergoes before its release to market. This stage also includes all regulatory requirements a medical product must meet before it is approved.

The distribution stage refers to the actual real availability of a product when it is put on the market, both in terms of physical access and reasonable affordability. Either of these factors of availability can be influenced by forces both internal and external to the market.

The final application and use stage encompasses all use of nanotechnology-based treatments by either patients or medical professionals and takes into account the treatment's intended purpose, mode of use, and any potential repercussions or side effects.

20.2.3 Ethical Issues in the R&D Stage

The foremost ethical issue associated with the R&D stage of nanomedicines is that the risk analysis undergone by both clinician and subject is fundamentally incomplete. Nanomedicines are so new that even the researchers who study them do not know of all the risks the nanomaterials they are based on pose to participant, researcher, and manufacturer alike [13]. This represents barrier to the establishment of one of the fundamental pillars of ethical clinical research, informed consent [14]. This illustrates the dilemma that nanomedicine researchers are placed in. Further research into nanomedicines would allow researchers to better enumerate the risks associated with them; however, it is this initial lack of information that would prevent that research from being ethically conducted in the first place. With a moratorium on research on nanomedicine similar to the one placed on genetic research aimed at cloning, after Dolly's, the cloned sheep's birth, neither likely nor advisable [10]. Researchers must carefully navigate this ethical issue and special care must be taken

to inform participants and researchers of the known risks and the scope of the unknown risks associated with the materials being researched.

In order to fully appreciate the gravity of the problem posed by the incomplete risk analysis, the known risks involved with nanomedicines must first be understood. The aspect of nanomedicines that gives them their novel properties as well as such potential to be dangerous is their size. Although when used properly in a controlled environment, the minuscule size of their component nanoparticles is what gives nanomedicines such novel properties, outside of this environment or even simply after they served their primary purpose, these particles can become dangerous to research subjects. One of the first qualities of nanoparticles noted by researchers was their ability to be absorbed through the pulmonary system ([13]; Warheit, [15]). While this quality makes nanoparticles an ideal delivery mechanism for nanomedicine vaccines, it also has the potential to lend itself to bioaccumulation of nanoparticles in the patient's lungs or even other organs [4, 13]. Furthermore, while nanoparticles' ability to cross cell membranes and even the blood–brain barrier makes them promising delivery methods for less damaging cancer treatments, it also allows nanoparticles to enter cells and parts of cells they are not designed to, exponentially increasing their health risk [16, 17]. However, it must be stated that these examples are hardly dispositive of the health risks associated with nanomedicines and researchers are only recently beginning to understand how they interact with the body's immune responses, much less any long-term side effects or toxic interactions ([13]; Resnik and Tinkle, [18]).

With such a dearth of understanding when it comes to the risks it is easy to see how nanomedicine R&D fails to constitute ethical research. Two of the seven requirements for an ethical clinical study are a favorable risk benefit ratio and the informed consent of the participant [14]. However, both of these requirements are dependent upon one of the researchers knowing the risks involved in the study, being able to perform an analysis on the risk benefit ratio, and relaying this information accurately to the participant. A researcher's ability to analyze the risk benefit ratio, whether it weighs societal or personal benefit against personal risk, is dependent upon the comprehensiveness of the information regarding risks they have access to, as any unaccounted for externalities might result in an erroneous and possibly dangerous finding of favorability. This lack of knowledge also undercuts the researcher's ability to convey these risks to the participant. Informed consent is generally regarded to have four requirements: disclosure, understanding, voluntariness, and competence [19]. Three of these requirements are undermined by a researcher's inability to comprehensively inform the participant of the risks. There can be no true understanding on the part of the patient if they, despite a researcher's best efforts, are not fully informed of the risks; a trial participant cannot truly understand the risks if they are not told them in the first place; and finally the participant's voluntary participation in the study can hardly be regarded as valid as it was established under a paucity of information regarding potential risks. Without meeting these requirements for ethical clinical research, nanomedicine R&D constitutes the unethical exposure of all involved to risks they are not even aware of.

An aside from this main issue is the fact that because nanomedicine is still very much an emerging technology that has not reached any semblance of technical

maturity yet; much of the R&D even when done on nearly finished nanomedicines is still largely very much research rather than actual therapy. This distinction is more than simply a semantic one, as participants and patients view the decision to participate in research as one that balances personal risk against societal benefit, while the decision to undergo treatment is one that balances personal risk against personal benefit (Resnik and Tinkle, [18]). A trial participant's misconception of this might skew their personal risk assessment, which compounding with existing ambiguities due to nanomedicine's untested nature, might lead them to make a decision they would not have otherwise (Resnik and Tinkle, [18]). Furthermore, it can be argued that because the trials required by regulatory agencies do not last longer than a few months, any long-term side effects of nanomedicines that could take years to manifest would not appear until such treatments have been used for decades (Resnik and Tinkle, [18]). This would effectively render all forms of nanomedicine treatment for the first few decades a large-scale research study rather than true treatment.

20.2.4 Ethical Issues in the Distribution Stage

The nature of scarcity in regards to any resource is that there will be those with it and those without. Upon its maturity, nanomedicine promises to be a fantastic resource, a boon to a nation's health care and overall quality of life. However, this will surely come at great cost. Like any new treatment, nanomedicines are and will likely continue to be for the forseeable future prohibitively expensive for developing countries [20]. But what burden does this place on the developed world to share its nanomedicine with the developing world? Questions of inequity of distribution and inequity of need weigh heavily upon society, as nanomedicines have the potential to save millions of lives in the developing world and some scholars warn that the created "nanodivide" between those with nanotechnology and by extension nanomedicine and those without could be a frictional point leading to war [21].

The part of this issue that makes the inequity of access so alarming is the inequity of need that serves as its backdrop. What makes the developing world's need for nanomedicine so tragic is the fact that its need for nanomedicine is so great. The sixth UN Millennium Development Goal to "combat HIV/AIDS, Malaria, and other diseases" underscores this inequity of need [22]. The World Health Organization (WHO) reported that in 2008 approximately 173 million people suffered from malaria, approximately 2.6 million people suffered from tuberculosis (TB), and roughly 5% of African adults suffered from HIV/AIDS [23]. Furthermore, the chance of dying before the age of 5 in the developing world is roughly 10%, 20 times that of the half a percent chance of dying before the age of 5 in the developed world [23]. Unlike the developed world, people in developing countries oftentimes do not have access to clean drinking water, a reliable cold chain, or regular professional medical care (Shetty, [4]). Traditional medicine is failing the developing world.

The major focus in nanomedicine is currently on early cancer detection through improved screening methods and more targeted chemotherapy delivery systems (Shetty, [4]). But nanomedicine's use in this area does not preclude it from promise in other areas as well. Nanotechnology-based labs-on-a-chip are used as cheaper and

faster alternatives to traditional TB tests and polymer nanoparticles or nanocapsules are theorized to make more effective and stable TB drug delivery systems [4]. Furthermore, these utilizations of nanotechnology could make aerosolized vaccinations for infectious diseases like TB possible [4]. These nanotechnology-based vaccines would have longer shelf lives and more importantly not require administration by a trained professional to be effective [4]. Finally, an Australian company created a nanoscale dendrimer gel, a synthetic polymer that interacts with cells, that prevents the spread of HIV/AIDS during sexual intercourse [24]. Found to be safe and effective, this dendrimer gel works by binding with proteins on the surface of targeted viruses and thereby preventing their attachment to human cells [24, 25]. In short, it is clear that despite nanomedicine's promise for developing countries, the likely focus in their development will be those applications that the developed world is interested in paying for.

The ethical issue here is one of the just distribution of scarce resources. However, the practical problem preventing effective and revolutionary nanomedicines from reaching those who need it the most is threefold. First, as the focus on anticancer treatments, a disease that is primarily a concern of the developed world, shows nanomedicines for the developed world are unlikely to be developed. Second, due to the high development costs of nanomedicines, they are likely to be too expensive for developing countries or even some countries with universal health care to afford [26, 27]. Finally, given their complexity and IP protections from first movers in the developed world, it is unlikely the developing world will be able to fabricate their own nanomedicines [28]. The grim outlook for the developing world in regards to nanomedicine is therefore as follows: they will not be able to afford nanomedicines they so desperately need and that they will not be able to make their own. One of the main questions that nanomedicine confronts society as a whole with is whether or not it will be provided to those who need it the most or if the world will "just accept a 'global nanodivide' as [it] has come to live with a 'global digital and genetic divide'" [11]. Perhaps, it is too much to hope for, but ideally the developed world would partner with the developing world and develop nanomedicines in conjunction with them. Similar to the stance Europe is taking, this would provide the developing world with life-saving treatments and while empowering them to take full advantage of this technology [21].

20.2.5 Ethical Issues in the Final Application and Use Stage

As with any treatment, nanomedicine as a medical treatment can be abused for enhancement purposes. While this dichotomy is not a new one and its application to nanomedicine is hardly surprising, nanomedicines' profound ability to impact the human body makes this issue of extreme importance. Some scholars draw the line between treatment and enhancement at the limits of "normal" human ability—this of course raises the question of normalcy, but that will be addressed in the following paragraph. To illustrate using an example, this line would classify a student with attention deficit hyperactivity disorder taking Ritalin to improve his or her failing grades to passing ones as treatment, but another student using the drug to get perfect

scores on exams to be enhancement [29]. When defined like this, this distinction is seemingly sensible and clear. However, people must draw this line for and focus groups have revealed that this treatment or enhancement debate is a concern held by most about nanomedicine [30].

The basis of the treatment or enhancement debate, however, is rooted in the concept of normalcy. But it is important to remember that the concept of what is "normal" is relative to a society. For instance, a five-foot-tall adult male would be abnormally short in the United States and a candidate for human growth hormone treatment, while they would be of normal height in a pygmy society where such administration of drugs could be viewed as an enhancement (Resnik and Tinkle, [31]). Thus whenever a relative qualifier is used, it is important to consider the particular circumstances in which it is used, especially for one as flexible as the term "normal." That being said, it is nearly impossible to make any broad rules that are nontrivial in their conclusions regarding what is and what is not normal and by extension what does and does not qualify as treatment as opposed to enhancement.

With the concept of what constitutes treatment and enhancement clarified, the ethical implications of a nanomedicine used for enhancement purposes can be discussed. In a vacuum, enhancement is not necessarily a bad thing. However, the problem arises when disparities exist within a society as they do in all real societies currently in existence. These disparities create inequality of access to the benefits of nanomedicine enhancements usually along wealth-based lines (Resnik and Tinkle, [31]). These wealth-based disparities would then translate into physical and physiological disparities with the wealthy possibly possessing immune resilience, strength, and even intelligence beyond normal human limits (Resnik and Tinkle, [31]). This disparity would then be exacerbated by relative advantages in job opportunities, academic, and athletic performance and in the extreme case, this could even lead to discrimination against those without nanomedicine enhancements (Resnik and Tinkle, [31]). However, it must be noted that not all schools of thought view these ethical issues negatively. While Christian scholars hold the traditional view and condemn this outcome, maintaining that people should embrace their imperfect selves [29]. Transhumanists believe that nanomedicine enhancements are an integral part in advancing humanity toward a technology enabled singularity at which, in a technology-enabled utopia, technology blended with the human body will enable to be whatever they wish to be [29].

While not directly related to the purpose of their use, widespread use of nanomedicines could also potentially lead to a level and type of environmental damage and widespread safety issues that researchers still fully do not understand. As discussion of the ethical issues of nanomedicine at the R&D stage showed, one of the major problems with the use of nanomedicines is the ambiguity that surrounds their research, manufacture, and use (Resnik and Tinkle, [31]). Many components of the nanodevices that will be used in nanotechnology-based treatments are toxic when not under the very specific circumstances that make them benign. For example, the gold nanoparticles that researchers use to image and destroy micrometastases are highly complex molecules and react differently under different circumstances [32]. Because of this varied behavior, the literature is divided on the toxicity of these

nanoparticles [33]. Another example is the often lauded carbon nanotube; these nanocomponents have the potential to replace gold and silver conductors in both traditional and nanoscale applications due to their high conductivity and relatively low cost [28]. These components will surely be fundamental for many nanotechnology-based treatments, but like gold nanoparticles have raised concerns about their toxicity and some have even compared their toxicological impact to that of asbestos [13, 34].

These nanoparticles are harmful not only to the individual using them but also potentially to the larger environment and the people they come into contact with. One of the primary risks associated with nanoparticles used in nanotechnology-based treatments is accumulation in the body—an accumulation of particles, however inert within the body can be just as deadly as the tumor they were first inserted to fight [32]. Therefore, in order to mitigate this risk, any benign nanomedicine's nanoparticles must have a means to eventually be consumed by or exit the body. This, however, necessarily would lead to the exposure of the environment to free nanoparticles. Because nanoparticles have a variety of vectors for entry into the human body, any exposure of them to the environment is potentially dangerous [33]. This problem is compounded by the fact that there is currently no proven method to recover free nanoparticles once released into the environment (Hoyt and Mason, [35]). Therefore, individuals using nanomedicines, their family members, and the public at large would all be at risk for contact with these free nanoparticles (Hoyt and Mason, [35]). At the same time, this might endanger the food supply as nanoparticles through biological magnifications could become concentrated in unsafe amounts in various animals humans consume.

Finally, while these cases should not be central in the discussion of nanoethics, extreme cases of the misuse of nanomedicines in the final application and use stage also pose grave ethical issues. Proponents of nanomedicine hope that later generations of nanotechnology-based treatments will be able to design macromolecules that would have the capability to interface with the nervous system at the tissue level [36]. When used responsibly, these capabilities could provide breakthroughs enabling better controlled prosthetic limbs, or prosthetic digits with a sense of touch. However, when used irresponsibly or maliciously, this could allow for the unauthorized access of one's thoughts, memories, or even mind. While restating the fact that this represents an extreme case that should hardly represent the character of this issue, when placed under greater scrutiny, these far-fetched alarmist concerns do not seem entirely improbable. If proponents of nanotechnology-based treatment readily accept the fact that one day nanotechnology would be able to translate one's thoughts into machine readable outputs and thereby enable the controlling of prosthetics, is it really that far of a stretch to have such a system write the data to memory instead. Once that basic premise is accepted, the question becomes one of access rather than of possibility and as discussed earlier in this book, personal medical devices' security are hardly robust (see Chapter 19). Scenarios like these raise the issues of personal autonomy and one's right to self-determination and despite the extreme nature of the circumstances, which would raise them, their gravity demands attention and vigilance against such unethical uses.

Unsurprisingly, the actual application and final use of nanomedicines gives rise to several ethical issues. These issues stem from both the evaluative question of how the nanomedicines are being used and the impact of their widespread use on the environment and human safety. Additionally, nanotechnology-based treatments raise other ethical issues in the extreme case in which they are intentionally used maliciously. The sources and scope of these issues demonstrate the fact that whenever nanomedicines are used, they will give rise to the issue of treatment or enhancement; when they are used by a large segment of society, environmental and safety ethical concerns might be raised; and when they are used maliciously, they may intrude on others' right to self-determination and individual autonomy. All of which demonstrates the need to be proactive in the discussion of the ethics surrounding nanomedicine so that this ethical discourse might guide the responsible development of nanomedicines and not lag behind the technology as Singer warns [8].

20.2.6 Ethical Narratives Used to Understand These Issues

The ethical issues raised at all stages in a nanotechnology-based treatment's life cycle are made much clearer and much more relatable when they are taken out of a vacuum and looked at through the lens of lay ethics. That is to say, one of the deficits of the institutional study of ethics is that discourse becomes bogged down by jargon and esoteric terms not readily understood by the public or scholars from other disciplines. Lay ethics, on the other hand, are shared and understood by all members of the public and despite their most vehement denials ethicists as well. This form of ethics is shaped by a society's historical and social experience and are commonly summarized not in terms like categorical imperative or deontology, but rather narratives or stories [37]. Generally speaking, when it comes to technology these narratives tend to express society's ambivalence regarding fundamentally new technologies, a trend that has been mirrored with nuclear technology, genetically modified crops, and other forms of biotechnology [38]. Furthermore, a common thread throughout these narratives is the underlying presumption of disaster of some kind. This presumption of catastrophe is possibly a type of self-defense mechanism of the human psyche preventing people from destroying themselves, by reminding them of a looming doom [39].

The first narrative expressed by the lay public is summarized by the theme "be careful what you wish for" [40]. This narrative acknowledges the promise of nanomedicines, but cautions that the promise they hold might not be what it seems or might be malignant rather than benign [40]. This encompasses major ethical themes and can easily be interpreted as paralleling much of the ethical issues identified in the application and final use stage. Viewing nanomedicine through this narrative lens raises questions such as the following: Would our healthcare system be ready to handle a rapidly aging society with death moratality due to cancer and other major causes of death nearly eliminated by nanomedicine?

The second narrative alluded to by the public was that of "Pandora's box" [40]. Advising caution in the face of an unknown but looming peril, this narrative is partly informed by the public's historical experience with crises like climate change or the asbestos crisis and partly by their imaginations fueled by sensationalist media [37].

The ethical issue of the dangerous biological magnification of nanoparticles derivative of nanomedicines fits neatly into this narrative.

The third narrative expressed by the public was the distaste for "messing with nature" [40]. This narrative is the fairly common notion that humans should not alter the natural order of the world and that some lines are simply not meant to be crossed [40]. Viewed as a means of achieving transhuman goals, nanomedicine is seen here as unholy or unnatural. This secular and religious notion is the basis for the treatment or enhancement debate at the center of several of the ethical issues that arise in the application and final use stage [29].

The fourth narrative was one of being "kept in the dark" [40]. This narrative reflects the feeling of powerlessness that the public feels despite efforts for firms and regulatory agencies to ensure transparency [40]. However, the root of this insecurity is far more likely the ethical issue of incomplete risk analysis identified in the R&D stage than any failure to report on the part of the government. Without complete knowledge of the risks of nanomedicine, the public will continue to view the lack of knowledge as evidence of its suppression [40].

The final narrative that the public expressed was that "the rich get richer" [40]. This narrative again acknowledges the positive impact that nanomedicine will have along with the corollary that these benefits will only be available to those, who can afford it [40]. This narrative is the embodiment of the ethical concept of distributive justice resonating with the lay public, and this directly reflects the ethical issue of inequalities of access and need identified in the distribution stage.

20.2.7 Are These Ethical Issues New?

Before the question of whether the ethical issues posed by nanoethics are truly new can be addressed, the definition of "new" must be clarified. Unlike its relatively simple meaning in common use, when used in the ethical sense, the meaning of "new" is more complex. "New" could refer to a novel occurrence of a previously recognized ethical issue or an occurrence of a novel ethical issue [41]. But what does it mean to be a new ethical issue? Van de Poel asserts that "an ethical issue is new if it is not raised by an existing technology or not dealt with in another field of applied ethics" [42]. While at first glance this definition seems satisfactory, an issue would be new if it is raised by a new technology, and it could also be seen as new if it has not been previously addressed. However, this first definition conflates the common use definition of "new" with the more refined one being used here. McGinn rightfully clarifies Van de Poel's original definition to be "an ethical issue is new if it is not raised by an existing technology and is not dealt with in another field of applied ethics" [43]. This clarification is extremely necessary to the practical value of this definition as an issue raised by a new technology could hardly be construed as new if it cleanly fits into an already established ethical framework. Thus, it is clear as McGinn points out that both parts of Van de Poel's original definition are needed simultaneously.

While clarifying how an ethical issue might be established as new, this discussion simply leads to the seemingly equally vexing problem of how to establish if an ethical issue is sufficiently dissimilar to existing ones such that it is not adequately

accommodated by existing ethical frameworks. While Van de Poel and McGinn's analyses stop short of this next question, Bacchini addresses it thoroughly. He somewhat circularly states that the sufficient and necessary condition for an ethical problem to be incompatible with existing fields of applied ethics is that it must possess a feature that is not possessed by an existing ethical problem type, new, ethically relevant, consistently ethically relevant, and not too weakly ethically relevant [41]. This of course begs the question of what "new" means here as well. Leaving much to be desired in terms of clarity, this tautological definition can be rephrased as requiring that "every possible means of paraphrasing it as an old ethical problem is such that it leaves at least one group of ethically relevant arguments supporting certain ethical conclusions about it has no corresponding group of paraphrased arguments supporting the corresponding ethical conclusion about the old ethical problem" [41]. All of which is to say that an ethical problem that truly does not fit into any existing ethical framework would be subject to supporting—and presumably detracting—ethical arguments that would not affect a conclusion in any existing ethical framework be it general or applied.

Having made the distinction between new ethical problems and simply new instances of old ethical problems and having defined what it means to be "new" as well as what constitutes an established ethical framework's inability to suitably deal with an ethical issue, the ethical problems posed by nanomedicine can finally be addressed. Taking the common ethical themes from the ethical issues from each stage in the lifecycle of a nanotechnology-based treatment, it is shown that these issues are not new ethical issues, but rather newer embodiments of old ethical issues. The issue of incomplete risk analysis and the validity of trial participants' informed consent in the R&D stage is nothing new to medical researchers testing new treatments [28]. Despite the multitude of trial stages a drug or device goes through before it is even tested on humans, there will always be some ambiguity of risk prior to actual human trials since this is the case for all treatments, not just nanomedicine (Resnik and Tinkle, [18]). The issue of inequality of use and access in the distribution stage is simply a matter of distributive justice [10]. This issue is not even unique to medical treatments in general but rather all resource inequalities. Finally, the issues in the application and final use stage all revolve around whether or not nanomedicines should be used and for what purpose thus alluding to the traditional therapy or enhancement debate [10]. These similarities are in fact so common that some scholars regard nanoethics as a form of new and emerging science and technology ethics or NESTethics [44]. Regardless of the validity or utility of this additional classification, Swierstra and Rip's assertion underlines the point that Grunwald, Bacchini, Allhoff, and many more make, that the ethical issues posed by nanotechnology and by extension nanomedicine are not new ethical issues but rather old ones repackaged in a new technology [10, 41, 45].

20.2.8 Nanoethics, Even When Applied to Nanomedicine Is Not Special. So What?

The parallels between nanoethics and the ethical discourse surrounding many other forms of new and emerging science and technologies do not discount its value as a field or lessen the need to discuss the issues surrounding nanotechnology and

nanomedicine [10, 44]. Therefore, while the creation of the field of nanoethics might be a pragmatic one, Godman stated it best that the issue of uniqueness of the field of nanoethics should not overshadow the importance of the questions it addresses [10, 46]. Furthermore, while the ethical questions raised might not necessarily be unique, the modes in which nanotechnology and nanomedicine raise them certainly is. As nanomedicine matures as a field, it will undoubtedly accomplish things that could not have even been dreamt before, which in turn will require the reassessment of how risk, hope, harm, and benefit are evaluated to accommodate these situations [41]. However, the ethical issues surrounding nanomedicine cannot be addressed in a strict purely utilitarian framework and the ethical framework used must incorporate principles of justice, autonomy, and personal integrity [28].

20.3 PRIVACY ISSUES

In the present day, privacy is a vanishing commodity. Personal privacy is routinely traded for convenience and access to the myriad of devices and databases people interact with daily. However, privacy was not always so readily discarded nor does it determine its current conception. It would seem that this cultural drift is both caused by and causing society's greater adoption and reliance on technology. Nanotechnology is predicted to exacerbate this trend [3]. With nanotechnology still coming into its own as a technology, nanoscale cameras and microphones still remain in the realm of science fiction for the time being. Because of this, whatever impact nanotechnology may have on privacy is speculative. But this is not to say that speculation about this impact needs to be uninformed. Several useful parallels can be drawn between current technologies and the predicted applications of nanotechnology in order to evaluate what impact those nanotechnologies have on one's privacy [3]. Once this broad picture is established, the analysis of nanotechnology can be focused on the specific implications of nanomedicine.

Like most technologies' benefits, the qualities of nanomedicine come at a cost. One of nanomedicine's main sources of promise is in its ability to analyze genetic material from within the body itself, which is hoped to lead to the proliferation of affordable personalized medicine [47]. However, the ultimate irony of this promise is that the greater the benefit of this personalized nanomedicine, the greater the threat it represents to individual privacy. Alternatively, when nanodevices are considered, these powerful, noninvasive, and nearly imperceptible devices will make medical data available to patients, medical professionals, and unauthorized individuals in an unprecedented way. Finally, whether it is personalized nanomedicine or nanodevices considered, both treatments will almost undoubtedly be integrated into e-healthcare systems for the storage of one's electronic healthcare records (EHRs) and thus cause even greater privacy concerns ([3], see Chapter 19). These privacy issues are particularly problematic due to the extremely revealing nature of the data at risk. While genetic information or personal health data as opposed to comprehensive medical records might not be as immediately interesting as one's internet browser history or emails, if used improperly they could fuel genetic discrimination.

In short, while the threat nanotechnology and nanomedicine pose to individual privacy does not necessarily prove fatal to the nascent technology, the value of privacy must be acknowledged and the danger to privacy that nanotechnology nanomedicine poses must be fully appreciated in its historical and legal contexts. Because only after the costs are accounted for can the benefit be appraised in light of it and only after the costs are understood can steps be taken to minimize them. These privacy issues affect all and should be discussed in the broadest venue and consist of a dialogue between all relevant actors [47].

20.3.1 Historical and Modern Conceptions and Treatments of Privacy

Like most of the principles that guide society, the current conception of privacy is very different from what it has been historically. The historical conception of privacy is perhaps best summarized by US Supreme Court Justice Louis Brandeis when he wrote in *Olmstead v. United States* "the right to be left alone is the most comprehensive of rights and the right most valued by civilized men" [48]. This historical conception of privacy finding its roots in the Post Office Act of 1710 and the Bill of Rights, extended to the person and their intimate possessions [49]. The text of the fourth amendment makes this scope—if not the right to privacy, as will be discussed during the legal discussion of privacy—in its text that reads "the right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures [50]. This concept makes immediate sense intuitively as an invasion of one's house or papers could be just as personally revealing as a search of one's person.

The historical conception of privacy is immediately seen to be limited in scope when compared to its modern counterpart. The modern conception of privacy appends to its historical definition, sentiment along the lines of "exercising [one's] right to decide for [oneself] whom [one] choose[s] to share the personal details of [one's] life with" [51]. This definition expands upon the historical one in two main ways. First, it broadens what is considered private from beyond one's person and intimate possessions. Second, it protects personal information after its initial collection. In addition to one's person, house, and papers, the modern conception of privacy seemingly adds any and all information that could be considered personal. Notable items within this list, but by no means dispositive, are credit ratings, DNA profiles, and phone records [50]. Given these few examples, it is an easy task to come up with dozens of others. This broadness in scope underlines one of the realities of our modern society: any information system could be a method of surveillance and an invasion of one's privacy [52]. Furthermore, this broadened definition even allows for more information on an individual to be collected. However, its second impact is arguably just as if not more important than its first: the choice of whom to share the personal details of one's life does not only limit its jurisdiction to the initial sharing of personal information, but also to the distribution of that information to third parties after the initial sharing. Therefore, unless authorized the modern conception of privacy allows a person to decide exactly who has access to his or her personal information, regardless of whom else it is shared with.

It is easy to see how the society it operates within changes a conception of privacy. At the time the Declaration of Independence was written, invasions of one's home and seizures of oneself and one's papers were resented. Thus, privacy to that society centered on protecting the person and their intimate possessions. However as technology and large information systems have taken ever larger roles in society, the modern conception of privacy has adapted information as an addition to this idea. In response to the recording and use of people's personal information, society realized the importance of this data and the methods of its seizure from individuals. However, it must be noted that not all cultures share these conceptions of privacy and what might constitute an invasion of privacy in one might be perfectly acceptable to another. To use the United States and the United Kingdom as an example, post 9/11 security measures were met with very different receptions. In the United States, these measures were met with resistance and public outcry, whereas in the United Kingdom nearly citywide closed-circuit television (CCTV) cameras in some cities is accepted as a part of life and not viewed as an "Orwellian intrusion" [53]. Counterintuitively, the European Union has far more comprehensive privacy protection of personal healthcare records than the United States providing a point of particular importance in the discussion of nanomedicine's impact on privacy (see Chapter 19). This fluidity of the concept is one of the difficulties faced when discussing privacy and for purposes of clarity this chapter will primarily address the issue in the United States.

20.3.2 Discussion and Review of Privacy Law

The cultural drift of the conception of privacy is minuscule compared to its legal evolution in the Untied States. The concept of privacy is so engrained in society that it is nearly impossible to imagine life without it. Legally, privacy, the rights it confers, and one's reasonable expectation of it are all central to much of the legal discourse that occurs today. However, the right to privacy as it is known today was not always such a well-recognized legal concept or even a right [54]. It was not until the Warren Court in *Griswold v. Connecticut* recognized the right to privacy as an unenumerated right, hidden but obvious within the penumbras of "specific guarantees in the Bill of Rights," that the right to privacy became a recognized piece of judicial precedent [55]. Prior to that point, one's right to privacy both interpersonal and from the government was stitched together in a patchwork of specific legislation and common law principles [56]. However, despite its recognition by the Warren Court, there is still significant discussion surrounding what constitutes a reasonable expectation of privacy especially in where it intersects with the fourth amendment's prohibition against unreasonable searches and seizures by the police. This contention is a testament to the importance of the right to privacy in everyday life, the importance it will have to nanotechnology and nanomedicine, and how ill-defined this concept still is.

Prior to the right to privacy's recognition by the Warren Court, privacy rights were created piece by piece through constitutional principles, specific legislation, and common law. For example, the third amendment specifically barred the quartering of soldiers in one's home, thus protecting one from that specific type of an invasion of

privacy and arguments can be made for soldiers surveillance function [57]. Likewise, the Post Office Act of 1710 preserved the privacy of one's mail by barring the government from reading it [49]. However, the common law origins of the right to privacy are less obvious. Supreme Court Justice-to-be Louis Brandeis at Harvard Law School with classmate Samuel Warren described this relationship in their seminal article "The Right to Privacy" [56]. They find that privacy is so broad a concept that it draws influence and influences contracts, IP, and libel [56]. However, none of these are perfect fits as it cannot be argued that there is a contract implied or explicit between two persons to preserve one's privacy; IP cannot be claimed because the harm of the breach of privacy is not dependent on the value of the information revealed, and the issue with the breach of privacy is not with the accuracy of the information revealed, but of the breach itself [56]. These imperfect attempts to derive the right to privacy from other areas of the law can be interpreted in one of two ways. It can be viewed as meaning the right to privacy does not exist as a singular concept and protection must be established piecemeal. Alternatively, it can be viewed as proving the right to privacy's separate existence as it could not be established through other means.

In *Griswold v. Connecticut*, the Warren Court decisively answered that question. Despite Connecticut law making it illegal to use, provide, or advise any medical treatment or medication for the purposes of preventing conception, Griswold the head of the Planned Parenthood League opened a clinic in Connecticut for that illicit purpose [58]. She was originally convicted under the law with the appellate court and the Connecticut Supreme Court upholding the conviction [58]. However, upon appeal to the Supreme Court, the Warren Court recognized the right to privacy, finding "specific guarantees in the Bill of Rights have penumbras, formed by emanations from those guarantees that help give them life and substance… various guarantees create zones of privacy" [55]. Specifically, these were elements of the first, third, fourth, fifth, and ninth amendments in the Bill of Rights [54]. In Griswold, it was this right to privacy that shielded the decision of married couples whether or not to use contraceptives from the influence of the law [55]. Justice Bork summarizes the opposing view in stating that "Justice Douglas… performed a miracle of transubstantiation… [not] disclos[ing]… how a series of specified rights combined to create a new and unspecified right" [59]. But regardless of how one views the Warren Court's opinion in Griswold, the fact remains that before Griswold the right to privacy did not exist as an explicitly recognized right, but after Griswold it became a tenable piece of judicial precedent.

Once recognized, the right to privacy and the reasonable expectation of privacy became central to fourth amendment challenges in the court. It is a general constitutional principle that illegally obtained evidence cannot be used at trial, this exclusionary rule is of obvious importance as Constitutional prohibitions against certain police tactics would be toothless if their fruits could still yield convictions. Thus, the legality of police surveillance tactics when they constitute a search due to the presence or lack thereof of a reasonable expectation of privacy is an extremely important question. One area of particular interest is the court's treatment of wiretaps and other forms of electronic surveillance, an issue with extreme

relevance to the development and future use of nanotechnology. In the 1928 case *Olmstead v. United States*, the Supreme Court found that the wiretapping of a public phone booth did not constitute a search [60]. Perhaps betraying the Court's dated analysis, Taft cited in his opinion on how the telephone wires passed through public space and thus were no longer protected by a reasonable expectation of privacy [60]. Fortunately, this precedent was overturned in the 1967 case *Katz v. United States*. Here the Court found that the entering of a telephone booth and shutting of the door implies the expectation of privacy, as one takes those measures to ensure the privacy of one's call in the public space [61]. Because of this, the wiretap constituted an invasion of privacy and a search [61].

The aforementioned fourth amendment discussion underlines the importance of, and ambiguity surrounding, the establishment of the legal parameters of privacy during interactions with technology. Prophetically Warren and Brandeis in 1890 stated that it is imperative for the law to protect people from "the enterprising press or any in the possession of any modern device for recording or reproducing sounds" [56]. With the vulnerability for legal privacy that technology offers was known to some even in 1890, it is hard to see how in a day and age defined by its connectedness, the means of said connection are hardly considered private. In 1986, the Electronic Communication Privacy Act was passed which afforded more protection from a privacy standpoint to written or physically recorded statements than their electronic counterparts [62]. This law, outdated at the time, is now nearly obsolete when it is increasingly difficult and meaningless to discern hard copies from the soft [62]. Some scholars predict that determining the level of protection afforded to these electronic communications will prove to be the most important constitutional question of the twenty-first century and surely one with great impact on nanotechnology and nanomedicine [62].

20.3.3 Impact of Nanotechnology in General

Before the privacy issues surrounding nanomedicine can be analyzed, the impact of nanotechnology in general on privacy must first be discussed. This is due in large part to the fact that nanomedicine—as well as nanotechnology for that matter—is still in its infancy as a field, and it is simply not known what these new technologies will look like or how they will function. By focusing on nanotechnology, in general, the outer parameters for the impact of nanomedicine can set, as intuitively the limits of nanomedicine are set by the limits of nanotechnology and not vice versa.

These outer limits can be set by focusing on current technologies that are used similar to how nanotechnology might be used in the future. Although nanotechnology has the potential to be a revolutionary technology, many of its benefits will be realized in its evolutionary applications (Kosta and Bowman, [63]). Because of this, future applications of nanotechnology will by no means be entirely dissimilar for those of current applications of technology. And "when near-future consequences of nanotechnology constitute a continuation of social changes instigated by earlier technologies, [there is] benefit from situating nanotech[ology] in the context of the earlier change" [50]. Simply put, by drawing these parallels between the current and the

speculative, the future impact of nanotechnology on privacy can be preemptively evaluated in an intelligent way [3]. Two current technologies' impact of privacy will be discussed here, but these are by no means the only parallels which can be drawn. Global positioning systems' (GPS) tracking abilities and CCTV surveillance will be evaluated and used to make informed predictions about the privacy implications of their equivalent nanotechnologies.

20.3.4 Parallels to GPS Tracking

As nanotechnology matures, it is likely that it may replace GPSs or be integrated into GPS systems to improve their performance, presenting a likely significant impact on individual privacy. Current GPSes, like the ones found within the console of many vehicles use readings from the receipt of microwaves emitted from satellites to triangulate position and heading [3]. In addition to its primary compass-like function, GPSes also have onboard memory that allows them to store some data regarding its past and present whereabouts; other GPSes might even have the ability to transmit their current location to an external server rather than storing the data on the device itself [3]. This functionality gives GPSes a lesser known secondary purpose that is to monitor the location, heading, and even speed of the attached vehicle [3]. When analyzed, this data could be used to reconstruct the daily habits of individuals and even to follow them. This tracking function raises significant privacy issues as does the question of what happens to the data after being transmitted and these issues mirror many that nanotechnology might face.

Like nanotechnology, GPSes present a plethora of useful functions. They can be used to drive to unknown locations, locate stolen vehicles, and trace 911 calls [3]. However, they are also subject to abuse. A case study of a Connecticut rental car company demonstrates GPSes' ability to infringe on privacy. After informing customers that "vehicles driven in excess of seventy-nine miles per hour will be charged $150 fee per occurrence. All [rented] vehicles are GPS equipped," American Car Rental Inc. used the equipped GPSs to monitor customers' speeds and enforce their anti-speeding policy [64]. However, customers were never given definitions of "occurrence," "GPS," or how their speed would be monitored [3]. With the probable miniaturization of equivalent tracking technology, it is no stretch of the imagination to envision a situation in which nanotechnology might be similarly abused. Consumers could be tracked and possibly penalized for noncompliance with agreements without their knowledge of the means of enforcement or their being monitored. More perversely, such actors might sell customers' real-time location information to third parties, who may further misuse the information and infringe upon customers' privacy.

Unexpectedly, an individual's privacy might receive protection by the courts in this area from such tracking conducted by the police. In the 2012 case *United States v. Jones*, the Supreme Court unanimously ruled that the installation of a GPS on a suspect's vehicle and subsequent tracking without a warrant or his or her consent established an unconstitutional fourth amendment search by the police [65]. This finding was reached on the basis that one's vehicle is classified as their "personal effects"

and that the installation of the GPS was a "search" as it was understood at the time the fourth amendment was written [65]. Because of this latter point, the Court found that it did not need to consider whether a reasonable expectation of privacy existed as per Katz, despite arguments that movement on a public road were readily observable by all and thus failed to establish a reasonable expectation of privacy [65]. Hopefully, courts will extend this same logic to GPS's nanotechnology-enabled progeny. However, this hope might prove to be exceedingly helpful in light of the government's increasingly bold use of technology. With recent National Security Agency (NSA) revelations and continued state and federal police use of stingray technology, it is doubtful whether nanotechnologies security benefits will be forgone by the government in favor of preserving personal privacy.

Whether by actual miniaturized radio receivers similar to those in GPSs, radioactive nanoparticle, or environmental nanosensors, the privacy issues raised by GPSs are likely to be mirrored and possibly exacerbated by nanotechnology. Worryingly, nanotechnology might make tracking considerably easier to conduct and harder to detect. This has the potential to increase surreptitious tracking of consumers and members of the public by corporate and government entities respectively. While business entities might be stopped through specific anti-tracking legislation as seen in many states with GPSs, the government is far less likely to be dissuaded or rather be self-regulate [3]. Furthermore, piecemeal privacy legislation is limited in scope and only adds to the list of "idiosyncratic private sector actions" protected by privacy policy [66].

20.3.5 Parallels to CCTV Surveillance

Unlike GPS-like tracking, CCTV-like surveillance is not achievable through nanotechnology's fundamental methods like the use of radioactive nanoparticles for tracking, but rather requires the miniaturization of cameras. But this is not to say that this is impossible or even unlikely. In fact given nanotechnology's fundamentally evolutionary nature, its assistance in this miniaturization is perhaps the most predictable application (Kosta and Bowman, [63]). Nanoscale CCTV cameras would have several advantages over traditional CCTV cameras. Current CCTV cameras operate by filming a location of interest and transmitting the video to a television screen where it is monitored typically by a guard or to a server where the videos are stored. These videos are then selectively viewed and if required, are forwarded to law enforcement. Thus, the quality of the recorded video is extremely important to CCTV's use to law enforcement. Nanotechnology will surely enhance the quality of CCTV as any technology that can serve to miniaturize the camera technology could also serve to improve the quality of what is recorded. However, if misused, this nanotechnology-based improvement could mean a heightening of the privacy risk that CCTV cameras entails.

The utility of CCTV systems lie in their ability to be interpreted, and it is in this interpreting in addition to the actual passive observation that represents the risk to privacy they pose. CCTV systems are much more than the recording of video as its true value and vulnerabilities lie in the "constructing, maintaining,

and mobilizing of images" [67]. At their core, CCTV systems are more human systems than technological ones and require the connection of CCTV staff, managers, local residents, and law enforcement [67]. Nanotechnology alone will be unable to improve these auxiliary functions, upon which CCTV systems depend. Nanoscale cameras could mean one of two things or both, more cameras in more places, or cameras in more personal places—due to their lower chance of detection. The first possibility would in turn increase the amount of interpretation that occurs while viewing the recorded video [68]. Therefore, more people would be needed to interpret the larger amount of footage, meaning that recordings of one's person would be exposed to more people, one's privacy would be invaded. The second possibility presents a much more straightforward invasion of privacy. If allowed by their nigh undetectable nature, nanoscale cameras are placed in more personal public areas, greatly reducing one's privacy in public. The extreme case of this scenario is the idea that nanoscale cameras could be placed on a person himself or herself and one's every action could be monitored.

While the near-constant surveillance of one's every action outside of his or her home is clearly a violation of one's right to privacy in the colloquial sense, legally such an Orwellian future would be condoned or at the very least consistent with the fourth amendment. Generally speaking, the Court has held that one cannot have a reasonable expectation of privacy if one's actions take place in public—that is, outside of their home and in the open. Specific examples of things that would not be covered by a reasonable expectation of privacy include items thrown in the trash, areas readily perceivable from public airspace, and events that occur in open fields or barns [69–71]. Therefore, law enforcement's use of CCTV cameras to film those public spaces does not constitute a fourth amendment search as the recording that takes place occurs in public where one does not have a reasonable expectation of privacy. As long as surveillance stayed within those bounds, regardless of how nanotechnology miniaturizes CCTV cameras or improves their image quality, their indiscriminate recordings would still be within the bounds of the fourth amendment. However, the Court would be far less likely to protect such recordings if the second scenario in the previous paragraph was realized. Once inside one's home, the reasonable expectation of privacy would be firmly established and law enforcement would be outside the bounds of the fourth amendment to be able to legally record. The same would hold true of anything nanotechnology-enabled CCTV cameras could record that would normally be imperceptible to one's regular eyesight.

When used in the traditional way, nanotechnology-enabled CCTV cameras would not impact one's privacy a significant amount more than their traditional predecessors. This is in part due to the Court's distinction between the expectation of privacy one has inside of their home and out, and partly due to the fact that nanotechnology would not foreseeably give CCTV cameras any radically new functionality. However, with ever-more creative surveillance methods employed by law enforcement, one can come up with numerous scenarios where this line begins to blur, greatly increasing the impact of nanotechnology. What if nanotechnology-based cameras become so common place that society begins to expect their use on anything investigated by law enforcement? Would this change one's expectation of privacy within

their home? What if a person knowingly came into contact with a law enforcement officer while knowing of their use of such surveillance technology? What if one left their window open and left for work? It is likely the Court would stick to its traditional distinction of location-based expectations of privacy and preserve the home regardless of a technology's use. However, the ambiguities these questions give rise to demonstrate the possibility for the infringement upon one's right to privacy by nanotechnology enabled CCTV as well as the importance of the Court in preserving our rights.

20.3.6 Privacy Issues in Nanomedicine

With an understanding of how nanotechnology is likely to impact individual privacy, informed predications can begin to be made about nanomedicine. One of the most salient concerns people have about nanotechnology is the loss of personal privacy, and this concern is likely to be intensified as nanotechnology becomes intertwined with health care in nanomedicine [72]. These concerns are not merely public conservative sentiment or resistance to change but rather well grounded in the very promises nanomedicine makes. Currently, "nanoscale instruments using cantilevers, electrical grids, quantum dots, and other devices are already providing better diagnostic information about one's genetic constitution, with molecular precision" [73]. These enhanced capabilities are not limited to genetic applications; nanomedicine and nanosensors also promise better medicines, delivery paths, as well as screening methods [50]. But all of these depend on the nanotechnology-based treatment's ability to detect or interact with one's body in an unprecedented way. Especially in the case of nanosensors, these interactions with the body will almost necessarily result in the generation of data which would be highly unlikely to be discarded immediately after utilization. The far more likely case is that the generated genetic or other medical information is wirelessly sent to servers to be stored and possibly integrated into a larger healthcare database. Once nanotechnology-based treatments become common in health care, the sharing of one's personal EHR and possibly even genetic data might be transferred between multiple parties and stored in multiple locations, all of which represent a possible breach of personal privacy ([3], see Chapter 19).

If left unchecked, the proliferation and ubiquity of nanomedicine in health care threatens to create a situation in which one would be required to surrender an unacceptable amount of personal privacy to receive health care [50]. Such a situation would be obviously and immediately problematic as it would force patients into the impossible choice of weighing one of their individual rights against a physical need. However, even in such a situation, it remains unclear as to whether or not one would even be able to legitimately waive one's right to privacy as it is difficult to comprehend how their data will be used. If the validity of such a waiver is impossible to establish, then any storage or transmission of a patient's HER or genetic data beyond the immediate purpose of providing them with medical care would constitute a violation of his or her right to privacy. This point is particularly important due to the fact that healthcare providers work

closely with insurance companies. A common adage is the first question an attending hospital physician will ask an emergency room patient is if they have insurance. The unauthorized, disclosure of one's genetic information to insurance companies or employers could have significant repercussions for the patient in the form of higher rates or wrongful termination. The availability of this information furnished by nanotechnology-based treatments could make genetic discrimination a frightening reality.

20.3.7 Waiving One's Right to Privacy?

The waiving of one's right to privacy is something done on a regular basis in modern society and is usually done while checking the "I agree" box under a terms and service agreement. However, the waiver of one's right to privacy is a far more serious action than its mundane regularity would suggest. When one waives one's right to privacy, they allow their personal information gained in a procedure or treatment to be used in whatever terms they agree to. However, as the provider of treatment, the healthcare provider usually provides the privacy waiver as well. These waivers can be written in ways that may unduly violate a patient's privacy or allow for the use of the patient's information in unrelated ways. As a patient receiving medical treatment is compelled by his or her medical condition to agree to obtain treatment, such a waiver can hardly be valid. However, this situation, albeit presently hypothetical, is not far from occurring. All of the factors, which would further invalidate a patient's waiver, are amplified in the case of nanomedicine: the need of the patient could be more dire; his or her understanding of the disclosure could be weaker; and the breadth of the disclosure could be greater.

Akin to informed consent unless very specific terms are met, it is nearly impossible to accept someone's waiver of their right to privacy as legitimate and actionable. To revisit the aforementioned case of *American Car Rental Inc. v. Commissioner or Consumer Protection* that involved the Connecticut rental car company, the Court found that for the rental car company's speeding fine to be upheld regarding the understanding its customers' had, three specific conditions had to be met. First they would have to "clearly and conspicuously disclose… the amount of the speeding fee in writing in the rental agreement" [64]. Second, they would have to "clearly and conspicuously disclose in writing in the rental agreement the use of a GPS device for the purpose of tracking the rental vehicle's speed" [64]. Third, they would have to "obtain… the customer's consent to the use of the GPS device for tracking the speed of the rental vehicle and to the amount of the speeding fee in writing immediately next to the foregoing disclosures" [64]. The terms of service agreements previously mentioned are legally held to a similarly rigorous standard and thus of questionable legality [74]. The legal requirement to notify people about conditions within an agreement is consistent with the ethical principle of informed consent that holds one must understand the consequences of their actions to consent. However, when it comes to the type of privacy disclosures nanomedicine will likely require patients to waive, it is unlikely such an understanding can be attained.

Patients do not readily understand medical information, let alone genetic information gathered through nanotechnology and stored in e-healthcare systems. The simple fact of the matter is that "a large part of the population… does not think in terms of atoms, molecules, and nanometers" [50]. Furthermore, it is unlikely a patient would even be aware of the dangers posed by the manner that healthcare providers use their medical information. Even if they were aware of the parties their information would be disclosed to, it is unlikely that patients would understand the use of their information by third parties and more importantly, the way those uses could negatively impact them. The extent of a patient's lack of knowledge regarding their medical information, combined with the disparity in bargaining power the two parties have at the time of a privacy waiver, works to undermine such a waiver's legal as well as ethical validity. Thus any use of a patient's medical information obtained through nanotechnology-based treatments beyond their immediate treatment could be viewed as breaches of privacy unless healthcare providers take special care to inform patients and ensure they understand what nanotechnology-based treatment is being used, what medical information it will collect, how that information will be used, who it will be transferred to, and how they will use it.

20.3.8 Nanomedicine-Enabled Genetic Discrimination

Due to its impact, the analysis of privacy in nanomedicine returns to the accuracy and accessibility of personal medical information. One form this information could take could be one's complete genetic code. These circumstances lend themselves to instances of genetic discrimination. This is not to say that people's genetics will be tested at birth by nanosensors and subsequently placed into a caste system, but rather that business interests might exploit people's genetic information at their expense in order to profit. If allowed to be shared with third parties, sensitive genetic information might end up in the hands of insurance companies' actuaries [3]. These actuaries could then—and rightfully by their trade—adjust the insurance rates of those in possession of certain genes. This, however, would be a form of genetic discrimination. This type of information could also be used by insurance companies to screen potential clients for markers for certain chronic and expensive diseases like genetic markers that make one more prone to certain types of cancers. But even if not shared with insurance companies by healthcare providers, the mere existence of the data could still pose a threat. Employers or insurance companies could pressure individuals into voluntarily disclosing this sensitive information [75].

20.3.9 Vulnerabilities from the Integration of Nanomedicine into e-Healthcare Systems

The integration of nanomedicine into e-healthcare systems also constitutes a significant threat to one's privacy. All of the data collected by nanotechnology-based treatments will have to be stored in e-healthcare systems and transmitted to and from these databases. However as previously noted by Mtenzi in an earlier chapter, these databases are by no means secure (see Chapter 19). Mtenzi also notes earlier, it is

possible to retroactively identify people even after they have been de-identified and then names can be connected to highly personal information such as one's mental health history or treatment for sexually transmitted diseases (see Chapter 19). Furthermore, it is unlikely that change in this area will be compelled by legislation. A somewhat notorious fact in privacy discussions is that the United States has better privacy protections for video rental lists than for electronic personal information like online marketing data [66]. This is consistent with Mtenzi and others' analysis that finds Europe's privacy laws to be far more comprehensive than the United States' ([66], see Chapter 19). While the vulnerability presented by integration of treatment into e-healthcare systems is not unique to nanomedicine, the increased sensitivity of the data nanotechnology-based treatments can collectively make this an ever more pressing issue.

20.3.10 Shrinking One's Right to Privacy

Privacy is an increasingly rare commodity, and it should be preserved whenever possible. However, the threats nanotechnology and nanomedicine pose to it are significant. While these threats are not entirely new, the capabilities of nanotechnology exacerbates and threaten the sanctity of one's home, person, and even DNA. Due to weak American privacy legislation, people are not protected from such infringements by private actors. For public actors, the courts do an acceptable job of protecting people from such intrusions through the fourth amendment. However, the courts will only maintain their importance if they can remain relevant by changing their rulings with the times. But nanotechnology and nanomedicine do not need to only represent threats to privacy. Optimistically, the accuracy of nanosensors will provide consumers with new and improved diagnostic tools to treat illnesses [50]. In this way, nanotechnology in personal medical devices could do for many illnesses what the home pregnancy kit did for women [50]. As nanotechnology matures, it will undoubtedly also provide similar means to increase one's privacy as well.

20.4 IP ISSUES

As more is learned about nanomedicine and new nanotechnology-based treatments or modes of fabrication are discovered, inventors and discoverers will wish to protect and assume some form of ownership of their brainchildren. The protection and ownership they seek is the recognition of their IP. Possession of such IP bestows upon the owner the right to ask the courts for relief from infringements on one's IP or with some forms of IP the right to a government-protected temporary monopoly of the sale and use of one's invention. There are four main forms of IP—patents, copyrights, trademarks, and trade secrets, and they each vary in the type of IP they protect, the protections extended, and the duration of the protection [76]. The best type of IP protection depends on the type of IP.

Patents will be paid special attention to in this section. Patents are issued through the United States Patent and Trademark Office (USPTO) and are arguably the most comprehensive form of IP protection for inventions that are fabricated as opposed to written or created in others ways by one's mind. Contrary to what its critics have to say about the declining quality of patents, the fact that it still takes over 2 years to acquire one speaks to the generally high level of rigor that they are vetted with at the USPTO [77]. The difficult process of acquiring a patent is made even more difficult when the applied patent for concerns nanomedicine. These nanomedicine patents are particularly difficult on both the patent applicant and the patent examiners at the USPTO trying to review the patent at nearly every stage of the process [77]. Despite these difficulties, or possibly because of these difficulties, it is extremely important for nanomedicine companies or startups to obtain a patent as they are crucial to the commercialization of a product when trying to recoup the high cost of R&D and are likely to play key roles in determining what companies control this emerging field [77].

20.4.1 Types of IP

Patents, copyrights, trademarks, and trade secrets are all different forms of IP varying in their disclosure requirements, protected subject matter, and level of protection extended. Generally speaking, disclosure requirements are directly related to protection with trade secrets on one end of the spectrum and patents on the other [76]. While they may differ in exact method, the goal of all of these forms of IP protection is the same: to ensure innovators and inventors are rewarded, incentivizing them to continue to innovate [76].

20.4.2 Patents

A patent is the granting of property rights by the USPTO to the inventor of "any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof" [76]. There are three types of patents: design patents, utility patents, and plant patents [76]. The distinctions between these three types of patents are as follows: design patents protect how an invention looks, utility patents protect how an invention operates, and plant patents protect a new and asexually produced species of plant [76]. As a form of personal and intellectual property, patents can be bought, sold, inherited, licensed, and even taxed [76]. The patent itself grants its owner the exclusive right to use, make, sell, and import the patented invention through the United States for 20 years in the case of utility and plant patents and for 14 years in the case of design patents [76]. To be patented, an invention must meet a variety of criteria, the invention "must be novel"; the invention must be "nonobvious to a person with knowledge in the field related to the invention"; and "it must have utility" [77]. Furthermore, the patent itself must describe the invention sufficiently to establish possession at the time of the application: "it must enable a person with knowledge in the field related to the invention to make or carry out the invention";

"it must enable a person with knowledge in the field related to the invention to use the invention"; "it must be described in clear, unambiguous, and definite terms"; and "it must set forth the best mode of making and using the invention" known to the inventor at the time of filing [77].

20.4.3 Copyrights

Copyrights or the literal rights to copy are forms of IP protection provided to the authors of "original works of authorship" [76]. These original works can be literary, dramatic, musical, artistic, or in other forms of intellectual work [76]. Generally speaking, they last for 70 years past the author's death [76].

20.4.4 Trademarks

A trademark is the protected mark or brand of a trade. This mark or brand is defined as any word, name, symbol, device, combination thereof, used or intended to be used "in commerce to identify and distinguish the goods or services" [76]. Colloquially, these marks are logos, but can also be banners, signs, or even smells [76]. A trademark lasts for 10 years and can be renewed for another 10-year period [76].

20.4.5 Trade Secrets

While still a formal form of IP protection, a trade secret does not involve the disclosure of the IP in question to the government. A trade secret is "any information that provides economic value that is not in the public domain and that has been reasonably kept secret" [76]. These trade secrets are protected indefinitely as long as they are kept secret [76].

20.4.6 Why Nanomedicine Is Particularly Difficult to Protect

Due to a variety of factors, nanomedicines and nanotechnology-based treatments are particularly difficult to patent at the USPTO. The most fundamental and more revealing factor is the confusion and disagreement on the USPTO definition of nanotechnology [78]. The USPTO along with the FDA use the definition of nanotechnology proposed by the National Nanotechnology Initiative (NNI) [78]. The NNI defines nanotechnology as "the understanding and control of matter at dimensions of roughly one to one hundred nanometers, where unique phenomena enable novel applications" [10]. This overly rigid definition completely overlooks the fact that nanotechnology represents not only one specific type but rather a cluster of technologies each with different properties and sizes [78]. Since ideal nanomedicial drug properties can be achieved through the use of slightly larger molecules, the hundred nanometer threshold is meaningless due to the non-criticality of the upper bound of the one hundred nanometers [78]. The use of ambiguous and oftentimes incorrect definition resulted in the USPTO's flawed preliminary classification system for

nanotechnology, exacerbating the difficulties USPTO staff have of sufficiently checking for prior art before issuing a nanotechnology patent [77].

Nanomedicine patents are exceedingly difficult to research prior art for, presenting another hurdle to a nanomedicine patent. The search for prior art by USPTO patent examiners is an integral part of the patent application process and is essential in vetting a patent to ensure that it is truly new invention. However, due to the flawed preliminary classification system, many nanomedicines are excluded from the USPTO's classification of nanotechnologies and which also excludes comparisons to their contemporaries when checking for prior art [77]. Furthermore because the technology is extremely new, much of the prior art in the field would not be found within the USPTO's patent database but rather in journals or books [77]. A final impediment to patent examiners' search for prior art is that the patenting environment incentivizes writing patents so that they are not easily found, such as by excluding the word "nano" from a nanomedicine patent, while the business environment simultaneously incentivizes the use of the word "nano" in non-nanomedical patents to attract interest [77]. With these difficulties in mind, it can truly be said "the informational burdens on examiners are clearly heavy—even before the examiner engages in the heavy lifting of interpreting the prior art" [79].

Once the patent examiner begins the task of reviewing the pieces of prior art and the patent itself, nanomedicine patents still present various issues. First and foremost, nanotechnologies in general are often so new that there are simply not enough experiments or experimental designs to demonstrate that a specific invention works [80]. In other cases, a nanomedicine might work in the exact way one expects it to and its implementation on the nanoscale merely maximizes these qualities, for example, silver nanoparticles displaying antibacterial properties, which are maximized by the nanoparticles' high surface area to volume ratio [10]. These uses might appear obvious to a patent examiner, so applicants might need to draft patents in terms of the industrial processes that are used to synthesize the molecules [81]. However, both of those difficulties assume examination by a qualified examiner. As nanomedicine is a fundamentally interdisciplinary field, nanomedicine patents are better examined by a team of examiners from different fields [77]. This problem is compounded by the USPTO's refusal to seek outside expertise on the topic [77].

Because of the aforementioned difficulties, it may be easier to keep some nanotechnologies secret rather than to patent them. The downside to protecting a product as a trade secret is the fact the secrecy is the only protective measure. Therefore, reverse engineering would threaten to reveal them. However, the extent of this deficiency, is dependent on the product's susceptibility to being reverse engineered. Highly popular and extremely famous products like Coca Cola and Kentucky Fried Chicken's recipes are protected as trade secrets and yet no one has been able to rediscover them [76]. Nanotechnology is similar to this case, since reverse engineering nanotechnology, especially more complicated devices, is far more difficult than in other fields of engineering [80]. Thus it might be more beneficial to avoid public disclosure of a nanotechnology's fabrication method and protect

it as a trade secret, especially in light of the difficulties faced when trying to patent nanotechnology.

20.4.7 Why Patents Are Important for the Development of Nanomedicines

Despite the difficulties they pose, patents remain extremely important for the developers of nanomedicines. The primary purpose of a patent for most applicants is to obtain the government granted monopoly for the sale and use of the patented product. A patent can mean recouping the enormous R&D costs that went into the development of the nanomedicine, while failure to acquire one could mean failing in the market at the hands of generics [82]. Patents give startups credibility with investors, and as a general rule patents usually precede funding for these startups [77, 83]. Given this importance, it is not surprising that companies are spending more and more of their budgets on acquiring and defending patents [83].

In addition to the primary purpose of obtaining a monopoly of the sales and use of the patented product, nanomedicine patents are being aggressively pursued due to applicant's desire to acquire building block patents and thus gain control over large sectors of the emerging nanomedicine field [77]. A patent also grants the holder exclusive licensing rights. Thus if a company or individual owns the patent of an essential building block of nanotechnology, then that company or individual may charge all else who wish to use their patented building block [77]. The right patent could allow a company to set up a licensing toll booth for the entire industry [77]. In cases like these, the social utility of patents becomes murky. If rightfully granted, the patent holder of a fundamental building block of an emerging technology should be rewarded, benefiting innovation. However, if that holder uses their patent to impede progress by charging exorbitant licensing fees or the patent was erroneously too broadly granted, innovation is harmed. Regardless of their value, it is clear that despite the system's flaws and the difficulty in obtaining them, patents are extremely important for nanomedicine's development and society as a whole.

20.4.8 Other Legal Issues

At this point in the chapter, it should be clear that with its many promises for a better world nanomedicine brings a variety of ethical, privacy, and IP issues. Ranging from the validity of the informed consent of nanomedicine trial participants to the possibility of the unauthorized use of one's genetic information, these issues are too important to be ignored and must be discussed in a public way with all relevant actors present. Such public discussions are necessary due to the fact that people are largely ignorant of nanomedicine as an emerging technology and that even among those who are informed, opinions regarding nanomedicine's value vary widely [82]. Furthermore, the magnitude of some of the risks posed as well as their dispersed nature requires the public's presence in the decision-making process as the public would be bearing a large portion of the risks of nanomedicine. Once suitable solutions to these issues have been created through the rise of nanomedicine, the question of implementation and regulation can be addressed.

20.4.9 A Regulatory Solution

Regulation holds the most promise in terms of implementing effective controls over the emerging nanomedicine industry. While the courts may serve as a somewhat effective protection against infringements of rights by the government, they do not play the regulatory role for the private sector effectively. It is simply not their job to do so. Likewise, issue-specific legislation is oftentimes not comprehensive enough to have a large impact and its piecemeal nature leaves the field fragmented and disorganized. Regulation to deal with the complex issues technical as well as nontechnical issues that nanomedicine presents are best handled by federally created regulatory agencies like the FDA. However, there are some that argue the FDA is incapable of effectively regulating nanomedicine [5]. While this might be true, its veracity does not dictate the abandoning of the FDA and the creation of a new nanotechnology-specific regulatory body, but rather the changing of inadequate policies [5]. Further oversight could be achieved with the creation of a centralized nanotechnology review board, which could evaluate the risks posed by nanomedicine better than the local institutional review board [5]. Such a board would regulate the research in conjunction with the FDA rather than instead of it, thus providing institutional expertise to supplement the FDA's established regulatory practices. This new body could then assist regulators in deciding whether or not federal review should be mandated for nanomedicine research, an important question given the gravity of the issues nanomedicine raises [5].

For such a regulatory solution to be implemented, one thing is certain: current regulations will need to be overhauled. One example is the regulation designed to ensure that the safety of materials encountered in public or in the workplace mandates the creation of materials safety data sheets (MSDS) [11]. Presently, the use of MSDS sheets for nanomaterials' traditional counterparts is accepted for the nanomaterials as well [11]. This is extremely perplexing and possibly dangerous as the purpose of the MSDS sheets is to allow for the better evaluation of a material's dangerous chemical properties by providing its composition and therefore inferentially its properties, whereas the purpose of most nanomaterials is to reduce the size of traditional materials to the point where their fundamental properties begin to change. This danger is seen with carbon nanotubes' use of the MSDS for graphite, which is technically correct due to the fact that they are both carbon compounds, but misleading insofar as they exhibit radically different physical and chemical properties [11]. Another example is the ambiguity surrounding the FDA classification of nanomedicines. Drugs, medical devices, and biological agents are all regulated very differently by the FDA, and compelling arguments for a nanomedicine's placement in any one of the categories can be made [84]. Until such ambiguities are dealt with, there will be confusion as to how nanotechnology-based treatments should be regulated [84].

An additional type of question to consider is what type of overall regulation strategy should be employed. Regulation can only truly be effective if those being regulated are part of the regulation process as well. This form of distributive regulation similar to that which occurs in the information and communication industry has the benefit of effectiveness as the key actors double as regulators and thus have both

the political will and means to enforce their self-set regulations [66]. However, the downside of such a distributive approach is that corporations, in the case of nanomedicine would likely be the large pharmaceutical companies, will be left to regulate important individual rights like one's right to know, right to privacy, or even right to individual autonomy [66]. Such a strategy need not replace a federal regulatory body to be effective either, but rather would supplement the body by adding industry's own monitors and auditors to the regulatory agency's. However, while the benefits of such a strategy are tempting, the regulatory agency must be careful not to get too close to the companies it is supposed to be watching.

20.5 CONCLUSION

Ultimately, these issues will not be addressed in any comprehensive manner without cooperation between the public, researchers, pharmaceutical and biotechnology companies, and regulatory agencies. The ethical issues posed by nanomedicine are too important to be dismissed without significant discourse between researchers and the affected public. Privacy issues will only be adequately solved if they are addressed in a bottom-up fashion from the biotechnology companies and healthcare providers themselves instead of a piecemeal top-down legislative solution from regulators. If nanomedicine innovations are to be protected without preventing future progress, a dialogue between legislators and industries need to be established. Similarly, and more importantly, if nanomedicine is to be regulated to the extent that the potential for harm the technology possesses demands, the public, industry, and legislatures all must work in tandem to create a regulatory body capable of handling the unique challenges posed by nanomedicine. However, given the promise that nanomedicine holds, it is likely that society will be able to rise to this regulatory challenge, allowing nanomedicine will be able to provide society with the hopeful and healthful future that so many predict it will.

ACKNOWLEDGMENTS

The authors are especially grateful to Mike Schlesinger for his generous guidance and support throughout the chapter. The authors are also thankful to Srinija Chintha for her contribution and insightful discussion on legal implications of nanomedicine.

REFERENCES

[1] Wagner V, Dullaart A, Bock A-K, Zweck A. 2006. The emerging nanomedicine landscape. Nat Biotechnol. 24(10): 1211–1217.

[2] Sadrieh N. 2006. FDA consideration for the regulation of nanomaterial containing products. Available at: http://www.fda.gov/nanotechnology/NIST_meeting_Houston-01-06.ppt (accessed on January 8, 2016). Slides 17–18.

[3] Brewer J, Gurel O. 2009. Nanomedicine: Issues of privacy and informed consent. Nanotech Law Bus. 6: 45–57.

[4] Shetty P. 2011. Can developing countries use nanotechnology to improve health?. Appropr Technol. 38(1): 9–12.

[5] Wolf S, Jones C. 2011. Designing oversight for nanomedicine research in human subjects: Systematic analysis of exceptional oversight for emerging technologies. J Nanopart Res. 13(4): 1449–1465.

[6] Asay C, Askin J, Baldwin C, Bessen J, Bock J, Boldrin M, Burnstein M, Chin A, Cohen L, Cohen W. 2015. Letter to congress. Available at: http://www.scribd.com/doc/257448042/IP-Scholars-Letter-to-Congress (accessed on January 8, 2016).

[7] Davies J. 2008. Nanotechnology oversight: An agenda for the new administration. Available at: http://www.nanotechproject.org/process/assets/files/6709/pen13.pdf (accessed on January 8, 2016).

[8] Mnyusiwalla A, Daar A, Singer P. 2003. 'Mind the gap': Science and ethics in technology. Nanotechnology. 14(3): R9–R13.

[9] Litton P. 2007. "Nanoethics?" what's new?. Hastings Cent Rep. 37(1): 22–25.

[10] Allhoff F. 2007. On the autonomy and justification of nanoethics. NanoEthics. 1(3): 185–210.

[11] Baird D, Vogt T. 2004. Societal and ethical interactions with nanotechnology (SEIN): An introduction. Nanotechnol Law Bus J. 1(4): 101–107.

[12] Keiper A. 2007. Nanoethics as a discipline?. New Atlantis A J Technol Soc. 16(1): 55–67.

[13] Oberdörster G, Oberdörster E, Oberdörster J. 2005. Nanotoxicology: An emerging discipline evolving from the study of ultrafine particles. Environ Health Perspect. 113(7): 823–839.

[14] Emanuel E, Wendler D, Grady C. 2000. What makes clinical research ethical?. JAMA. 283(20): 2701–2711.

[15] Warheit, D. 2010. Assessing health risks of inhaled nano materials: Development of pulmonary bioassay hazard studies. Anal Bioanal Chem. 398: 607–612.

[16] Geiser M, Rothen-Rutishauser B, Kapp N, Schürch S, Kreyling W, Schulz H, Semmler M, Im Hof V, Heyder J, Gehr P. 2005. Ultrafine particles cross cellular membranes by nonphagocytic mechanisms in lungs and in cultured cells. Environ Health Perspect. 113(11): 1555–1560.

[17] Hoet P, Brüske-Hohlfeld I, Salata O. 2004. Nanoparticles—known and unknown health risks. J Nanobiotechnol. 2(12): 12–27.

[18] Resnik D, Tinkle S. 2007a. Ethical issues in clinical trials involving nanomedicine. Contemp Clin Trials. 28(4): 433–441.

[19] Shrader-Frechette K. 2007. Nanotoxicology and ethical conditions for informed consent. NanoEthics. 1(1): 47–56.

[20] Timmermans J, Zhao Y, Van de Hoven J. 2011. Ethics and nanopharmacy: Value sensitive design of new drugs. NanoEthics. 5(3): 269–283.

[21] Malsch I. 2013. Governing nanotechnology in a multi-stakeholder world. NanoEthics. 7(2): 161–172.

[22] United Nations. Goal 6: Combat HIV/AIDS, malaria, and other diseases. Available at: http://www.un.org/millenniumgoals/aids.shtml (accessed on January 8, 2016).

[23] World Health Organization. 2010. World Health Statistics 2010. Geneva: World Health Organization Press.

[24] Rupp R, Rosenthal S, Stanberry L. 2007. Vivagel (SPL7013 gel): A candidate dendrimer—microbicide for the prevention of HIV and HSV infection. Int J Nanomed. 2(4): 561–566.

[25] Abner S, Guenthner P, Guarner J, Hancock K, Cummins J Jr., Fink A, Gilmore T, Staley C, Ward A, Ali O, et al. 2005. A human colorectal explant culture to evaluate topical microbicides for the prevention of HIV infection. J Infect Dis. 192(9): 1545–1556.

[26] Bennett-Woods D. 2008. Nanotechnology: Ethics and Society. Boca Raton, FL: CRC Press.

[27] Mehta M. 2004. The future of nanomedicine looks promising, but only if we learn from the past. Health Law Rev. 13(1): 16–18.

[28] Kermisch C. 2012. Do new ethical issues arise at each stage of nanotechnology development?. NanoEthics. 6(1): 29–37.

[29] Toumey C. 2011. Seven religious reactions to nanotechnology. NanoEthics: 5(3): 251–267.

[30] Davies S, Kearnes M, Macnaghten P. 2009. 'All things weird and scary': Nanotechnology, theology and cultural resources. Cult Relig Interdiscip J. 10(2): 201–220.

[31] Resnik D, Tinkle S. 2007b. Ethics in nanomedicine. Nanomedicine. 2(3): 345–350.

[32] Fadeel B, Garcia-Bennett A. 2009. Better safe than sorry: Understanding the toxicological properties of inorganic nanoparticles manufactured for biomedical applications. Adv Drug Deliv Rev. 62(3): 362–374.

[33] Brayner R. 2008. The toxicological impact of nanoparticles. Nano Today. 3(1–2): 48–55.

[34] McGee E. 2009. Nanomedicine: Ethical concerns beyond diagnostics, drugs, and techniques. Am J Bioethics. 9(10): 14–15.

[35] Hoyt V, Mason, E. 2008. Natotech emergency health issues. J Chem Health Safety. 15(2): 10–15.

[36] Davies C. 2009. Oversight of next generation nanotechnology. Available at: http://www.nanotechproject.org/process/assets/files/7316/pen-18.pdf (accessed on January 8, 2016).

[37] Ferrari A, Nordmann A. 2010. Beyond conversation: Some lessons for nanoethics. NanoEthics. 4(2): 171–181.

[38] Renn O, Roco M. 2006. White paper on nanotechnology risk governance. Available at: http://www.irgc.org/IMG/pdf/IRGC_white_paper_2_PDF_final_version-2.pdf (accessed on January 8, 2016).

[39] Dupuy J-P. 2010. The narratology of lay ethics. NanoEthics. 4(2): 153–170.

[40] Davies S, Macnaghten P. 2010. Narratives of mastery and resistance: Lay ethics of nanotechnology. NanoEthics. 4(2): 141–151.

[41] Bacchini F. 2013. Is nanotechnology giving rise to new ethical problems?. NanoEthics. 7(2): 107–119.

[42] Van de Poel I. 2008. How should we do nanoethics? a network approach for discerning ethical issues in nanotechnology. NanoEthics. 2(1): 25–38.

[43] McGinn R. 2010. What's different, ethically, about nanotechnology? Foundational questions and answers. NanoEthics. 4(2): 115–128.

[44] Swierstra T, Rip A. 2007. Nano-ethics as NEST-ethics: Patterns of moral argumentation about new and emerging science and technology. NanoEthics 1(1): 3–20.

[45] Grunwald A. 2005. Nanotechnology—a new field of ethical inquiry? Sci Eng Ethics 11(2): 187–201.

[46] Godman M. 2008. But is it unique to nanotechnology?: Reframing nanoethics. Sci Eng Ethics. 14(3): 391–403.

[47] Ferrari M, Philibert M, Sanhai W. 2009. Nanomedicine and Society. Clin Pharacol Ther. 85(5): 466–467.

[48] Brandeis L. 1928. Olmstead v. United States. 277 U.S. Reports. 438, 478.

[49] Li J. 2003. The Center for Democracy, Technology, and Internet Privacy in the U.S.: Lessons of the Last Five Years. Lanham, MD: Scarecrow Press.

[50] Toumey C. 2007. Privacy in the shadow of nanotechnology. NanoEthics: 1(3): 211–222.

[51] Cavoukian A, Tapscott D. 1997. Who Knows: Safeguarding Your Privacy in a Networked World. New York, NY: McGraw-Hill Higher Education Textbook Publisher.

[52] Monahan T. 2006. Preface. In: Monahan T (ed.), Surveillance and Security. New York, NY: Routledge.

[53] Landler M. 2007. Where little is left outside the camera's eye. *New York Times*. July 8, 2007. WK5.

[54] Luban D. 1999. The warran court and the concept of a right. Harv Civ Rights Civ Liberties Law Rev. 34(1): 7–37.

[55] Douglas W. 1965. Griswold v. Connecticut. 381 U.S. Reports. 479.

[56] Warren S, Brandeis L. 1890. The right to privacy. Harvard Law Rev. 4(5): 193–220.

[57] Moran F. 1980. Free speech, the military, and the national interest. Air University Rev. 1980 (May–June).

[58] Carroll L. 2012. The Baby Matrix: Why Freeing our Minds from Outmoded Thinking About Parenthood and Reproduction Will Create a Better World. LiveTrue Books.

[59] Bork R. 1971. Neutral principles and some first amendment problems. Indiana Law J. 47(1): 8–9.

[60] Taft W. 1928. Olmstead v. United States. 277 U.S. Reports. 428.

[61] Stewart P. 1967. Katz v. United States. 389 U.S. Reports. 387.

[62] Waltz B. 2014. Privacy in the digital age. Indiana Law Rev. 48(1): 205–211.

[63] Kosta E, Bowman D. 2011. Teaching or tracking? Regulatory challenges of nano-enabled ICT implants Law Policy. 33(2): 256–275.

[64] Borden J. 2005. American Car Rental Inc. v. commissioner of consumer protection. 273. Conn Reports. 296.

[65] Scalia A. 2012. United States v. Jones. 565 U.S. Reports.

[66] Newman A. 2015. What the "right to be forgotten" means for privacy in the digital age. Science. 347(6221): 507–508.

[67] Neyland D. 2015. Closed circuits of interaction? Inf Commun Soc. 7(2): 252–271.

[68] Mehta M. 2002. Privacy vs. surveillance: How to avoid a nano-panoptic future. Can Chem News. 54(10): 31–33.

[69] Blackmun H. 1979. Smith v. Maryland. 442 U.S. Reports. 735.

[70] Burger W. 1986. California v. Ciraolo. 476 U.S. Reports. 207.

[71] Powell L. 1984. Oliver v. United States. 466 U.S. Reports. 170.

[72] Cobb M, Macoubrie J. 2004. Public perceptions about nanotechnology: Risks, benefits, and trust. J Nanopart Res. 6(4): 395–405.

[73] Alivasatos A. 2001. Less is more in medicine. Sci Am. 285(3): 66–73.

[74] Sack R. 2012. Schnabel v. trilegiant corp. and affinion Inc. 11–1311. (2nd Circuit).

[75] Brune H, Ernst H, Grunwald A, Grünwald W, Hofmann H, Kruh H, Janich P, Mayor M, Buxton D, Lee S, Wickline S, Ferrari M. 2003. Recommendations of the national heart, lung, and blood institute nanotechnology working group. Circulation. 108(22): 2737–24742.

[76] Purvis S. The fundamentals of intellectual property for the entrepreneur. Available at: http://www.uspto.gov/sites/default/files/about/offices/ous/121115.pdf (accessed on January 8, 2016). Slides 7–17.

[77] Bawa R. 2007. Patents and nanomedicine. Nanomedicine. 2(3): 351–374.

[78] Bawa R. 2005. Will the nanomedicine "patent land grab" thwart commercialization?. Nanomed Nanotechnol Biol Med. 1(4): 346–350.

[79] Petherbridge L. 2006. Positive examination. IDEA Intellect Property Law Rev. 46(2): 173–220.

[80] Zekos G. 2006. Nanotechnology and biotechnology patents. Int J Law Inf Technol. 14(3): 310–369.

[81] Kim K. 2008. Implications of recent U.S. supreme court IP rulings for nanomedicine patents. Nanomedicine. 3(2): 141–143.

[82] Bosetti R, Vereeck L. 2011. Future of nanomedicine: Obstacles and remedies. Nanomedicine. 6(4):747–755.

[83] Bawa R, Bawa S, Maebius S, Flynn T, Wei C. 2005. Protecting new ideas and inventions in nanomedicine with patents. Nanomed Nanotechnol Biol Med. 1(2): 150–158.

[84] Buxton D, Lee S, Wickline S, Ferrari M. 2003. Recommendations of the national heart, lung, and blood institute nanotechnology working group. Circulation. 108(22): 2737–2742.

PART V

CONCLUSION

21

CONCLUDING REMARKS

Zhaoqi Chen, Mary Mehrnoosh Eshaghian-Wilner, Kalyani Gonde, Kodiak Ravicz, Rakshith Saligram and Mike Schlesinger*

EE-Systems, University of Southern California, Los Angeles, CA, USA

21.1 WIRELESS COMPUTING IN HEALTH CARE

The new state of the art in medical research revolves around Wireless computing. Many of the devices and other equipment currently employed throughout modern hospitals are related to Wireless computing, and the medical devices of the future are not likely to change their dependence on that technology. The sheer amount of data produced by these modern facilities have led to diagnostic accuracy and efficiency issues, big data problems, and medical data security challenges. As Chapter 2 [2] details, recent research into pervasive computing and big data processing—specifically in nanometer electronic systems, data compression-based information processing, and cloud computing technologies—hold great potential to create new, novel platforms for new generations of medical sensing and monitoring systems. In other words, health care can be anywhere and everywhere. Patient monitoring technologies hold the potential to support continuous care, no matter where the patient is at any time.

In Chapter 1 [1], we introduce the concept of wireless computing in medicine, demonstrating the two primary ways in which wireless technology can make healthcare more effective. This trend toward Wireless health care is important in that it allows healthcare professionals to be more effective in their treatments in two distinct ways. The first and most obvious is instant care. If a monitored patient experiences an anomaly—for instance, a heart attack or an embolism—their doctors can be

*Authors are listed alphabetically by last name, as opposed to by order of contribution.

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.

immediately notified and take appropriate action right away. Wireless health care can also assist medical professionals through the gathering of mass amounts of data. The treatment and care of patients with chronic diseases, such as certain neurological, respiratory, and cardiovascular diseases, requires constant monitoring of the patient over a long period of time—sometimes over the entire life of the patient. This monitoring is currently carried out in person at a hospital, which is ideal for neither the patient, who must be physically present in the hospital for the entire lifetime as well as pay for the treatment, nor the hospital, which has limited space and resources. In Chapter 3 [3], we detail how pervasive computing can make health care in hospitals more effective and convenient. Beyond the hospital, ubiquitous computing, through wearable sensor technologies and smartphone applications, can make Wireless health care both available and attractive to the patients suffering from chronic disease. Chapter 4 [4] discusses platforms for chronic disease management, which includes a wireless body sensor network, which collects important data from the patient, and a mobile base unit, which receives data wirelessly from the wireless body sensor network and transmits it on. Monitored data includes blood pressure, glucose level, heart rate, ECG, blood oxygen saturation, breath rate, and other biomarkers throughout the body. These sensors—which can potentially come in the form of rings, watches, or even tattoos—drastically decrease the inconveniences of constant monitoring for both the patient and the hospital.

While independent wearable sensors are the near future of Wireless medicine, they are not fully available right now. Right now, we have to make use of technologies that everyone has with them the vast majority of the time and has the technological ability to both collect and transmit useful data. Mobile phones fit the bill almost perfectly. They can collect data on behavior, location, communications, and levels of physical activity, as well as other health data. These data are currently used in the field of behavioral epidemiology, which is the study of how lifestyle and behavior relate to the occurrence of a disease. Numerous diseases provide early evidence of their onset from changes in behavior, which can present long before the disease is medically diagnosed. The collection of data from mobile phones can provide key insight into patients' habits, and potentially show a pattern indicating the presence of one of these diseases. The analysis of mobile phone data has been the focus of recent attention in Wireless computing, giving rise to the field known as mobile sensing. In Chapter 5, the author presents InCense: a collaborative mobile sensing toolkit developed to facilitate behavioral data gathering from populations of mobile phone users. InCense is designed to provide people with little to no technical background with a tool that assists in the rapid design and implementation of mobile phone sensing campaigns. Two InCense-supported studies are described in Chapter 5 [5], geared specifically toward detecting frailty in the elderly and disruptive behaviors among people suffering from dementia. Through these studies and others like them, we could potentially uncover data that could be used to develop behavior-aware healthcare systems for applications that range from providing early diagnosis to uncovering lifestyle causes of mobility, inducing behavior change, helping in disease management, and dealing with problematic behaviors.

The ideas inherent to behavioral epidemiology—specifically the concept of monitoring patients' behavior and other data points—can be applied to other areas as

well. One specific area in which these concepts are applied is in assisting those with cognitive disabilities. The World Health Organization (WHO) indicates that 1–4% of the world population may have some level of cognitive disability—the presence of incomplete or arrested mental development, principally characterized by the deterioration of concrete functions at each stage of development and that contribute to the overall level of intelligence. People with cognitive disabilities can experience deficits in memory, executive function, motor impairments, and other cognitive processes such as learning and using language. These deficits can in turn lead to lower perceived quality of life and self-care abilities. Recently, Wireless computing has become a focal point in efforts to support the independence, therapy, and rehabilitation of individuals with cognitive disabilities. As discussed in Chapter 5 [5], wearable sensors could make the required collection of mood, emotional, and behavioral data much more convenient for both doctors and patients alike, as well as provide doctors with new information not previously acquirable through mere observation. In Chapter 6 [6], however, we see that for applications such as assisting those with cognitive disabilities, Wireless computing can do more than simply assist with data collection. In fact, it can help treat cognitive disabilities directly. Augmented reality and tangible computing, as well as the use of situated displays and mobile applications, can support cognitive therapies, provide step-by-step guidance, and assist nonverbal individuals with alternative methods of communication. Digital systems can also provide therapy through games to support engagement, the development of motor skills, and maintain the adherence of individuals with cognitive disabilities during motor and rehabilitation therapies. Wireless computing is especially active in the fight against Alzheimer's disease and childhood autism, so those two diseases are covered in depth in Chapter 6 [6].

Clearly, mobile sensing and Wireless devices are already playing a vital role in multiple fields of medicine, and their role will only grow more important as time passes. In the future, more devices will be implantable under the skin, similar to pacemakers today. Once placed within the body, these devices will be largely inaccessible to any method but surgery. This poses a problem: How do we power these devices? Wires are not an option; nor are batteries that must be surgically replaced with regularity. There is only one truly viable option to power these increasingly tiny devices: wireless power. In Chapter 7 [7], we introduce different methods of wireless power and discuss the potential for this technology in the realm of implantable medical devices. There are three main approaches to wireless power transfer: the electromagnetic wave approach, inductive coupling, and magnetic resonance coupling. While these techniques have in fact been proven to work, there are still limitations as to the efficiency, size, and cost of these wireless power transfer systems. These are due to the expense of the materials required, the tiny scale of the desired applications, and theoretical restrictions imposed by the laws of physics on those scales. There have been a number of advances in recent years, however, resulting in a number of industrial products and standards. These standards can also be applied to medical implantable devices, giving them the ability to operate freely throughout the body. This may well result in smaller, more space-efficient Wireless devices better equipped to accomplish their missions within the body.

While wireless power does hold the potential to provide all necessary power to an implanted system, we are not yet able to achieve that goal by those means. Also, it would be ideal to use the power that we do have as efficiently as possible. In order to maximize power efficiency and system lifespan, we have algorithms such as those presented in Chapters 8 and 9, this volume [8, 9]. KNOWME, a network of sensors deployed throughout the body, works by making decisions as to which sensors to listen to and when. For example, it does not make sense to listen to sensors in less vital areas of the body with the same frequency as those in more vital areas, such as the heart or lungs. If we can divert power and resources from the non-vital sensors to those in more vital areas, we can poll those vital sensors more often and end up with more useful data overall. The solution presented in Chapter 9 [9] addresses similar concerns as KNOWME. The technique used here enables distributed sensor nodes to change their sampling rates in response to changing conditions at each node. To accomplish this, a stochastic Markov decision process (MDP) model is used to dictate sampling rates for each node across the network in all possible states. Because all of the actual computations are done offline, the system must only look up the appropriate next stage for each node rather than computing it in real time. This, along with other decisions made in designing the process, allow the entire system to function with greater efficiency and less downtime over a longer total system lifespan. Given these benefits over an unregulated system, MDP models such as KNOWME and the one presented here could be invaluable when deploying sensor networks of any real size and importance throughout the body.

21.2 NANOMEDICINE

In the first part of this book, we explored the field of Wireless computing. In the second, we investigated the microscopic world of nanomedicine. Nanotechnology can be applied to the field of medicine in numerous ways. In the following sections, we will recap many of the techniques and technologies detailed in previous chapters.

This section deals specifically with medical nanotechnology, or nanomedicine. Chapter 10 [10] serves as an introduction to this field. Nanomedicine has the potential to fundamentally change the way we treat disease. As applied specifically to cancer, nanomedicine has two major benefits over traditional treatments: concentration and precision. Standard cancer treatments release a large amount of drugs into the general area of the disease. These drugs do not differentiate between cancerous and healthy cells—they kill them all. Nanomedicine, however, has the ability to target cancerous cells on the cellular level. This remarkable precision allows the drug-carrying nanorobots to avoid damaging healthy cells, which vastly improves the efficiency of the treatment while also eradicating many of the potentially deadly side effects of today's treatments. Due to their small size, we are able to employ many of these nanorobots at once in large concentrations, theoretically resulting in an effective, fast-acting treatment. There are many unsolved questions still to be answered, however. Challenges such as nanoscale biocompatibility, system power, and

nanorobot maneuverability—among others—still remain between us and our final goal of a nanorobot-based cure for cancer. The potential, however, is there—we have to work to make it a reality.

While these nanorobots may one day have the type of enormous impact on medical treatment that scientists have long anticipated, we still have many challenges to overcome before we can begin to save lives. We are, however, making progress. One group in particular has recently made enormous strides in the detection and treatment of cancer at the cellular level through the use of magnetoelectric nanoparticles (MENs). MENs are very similar to the more common magnetic nanoparticles (MNs), but they have a few important differences. Like MNs, MENs can be guided throughout the body by the application of an external electric field. This gives doctors the ability to manually direct them to the affected area of the body, resulting in a higher efficiency rate as compared to unguided nanoparticles. Unlike MNs, however, MENs do not dissipate heat as a result of this guidance maneuver. Heat from MNs can potentially do harm to the body, so this quality of MENs is beneficial. MENs are preferable to MNs due to their unique abilities in drug delivery as well. In order to release the attached medication, MNs require either heat or mechanical pressure. MEN, however, can release their medication due to the application of specific magnetic fields. This does not expose the body to either heat or mechanical pressure, increasing both safety and energy efficiency. In addition, MENs can be used to produce electric fields near cancerous cells. These fields can change the size of the pores in the membrane of the cancerous cell, greatly increasing drug intake efficiency. These differences are significant enough to give MENs a distinctive advantage over MNs as drug delivery units. As described in Chapter 11 [11], MENs have already been proven effective against cancer cells. When utilized against an ovarian cancer tumor, MENs were able to deliver a significantly higher dosage of medication through the membrane of the cancerous cells as compared to MNs or other delivery methods. As a result, the tumor was effectively eradicated within 24 h. This trial, as well as a second trial also described in Chapter 11 [11] involving the delivery of anti-HIV medication, displays the power that MENs have in the fight against disease. As more research is conducted, MENs will be seen as particularly promising options for drug delivery.

Inorganic nanorobots and nanoparticles may well be the path to the eventual cure for cancer, but there are other options as well. In Chapter 12 [12], we explore DNA computing as one such option. DNA computing is a method that maintains the theory of standard inorganic computing, but accomplishes its goal through organic means. DNA computing uses the four common DNA base pairs to form standard logic gates that can then be arranged into functioning electronic circuits. Scientists can even theoretically use these base pairs to code for self-assembling nanoscale structures, similar to the inorganic nanorobots described earlier. This would be a major development, as self-assembling nanoscale structures would remove the need to manually fabricate nanorobots at such a small scale. Also because they are made of organic material that originated inside the body, it may be easier to make them fully biocompatible within the human body. If organic nanoscale structures are able to overcome this problem so easily, they have a major advantage over their inorganic competition.

While Chapter 11 [11] provides a detailed look into one particularly promising method of detecting and treating cancer cells, that is not by any means the only—or even necessarily the best—solution. There are countless solutions currently being researched, and likely many more not yet discovered. Chapter 13 [13] discusses graphene—an allotrope of carbon—that can be used in creating biosensors. Due to a number of intrinsic traits, graphene can be tuned to sense a wide range of proteins in the bloodstream. This can include proteins produced by cancer cells, making graphene a relatively cheap, eco-friendly, and efficient way to detect cancer cells within the body. Once the cancerous cells have been detected, any one of an array of previously designed termination methods can be employed to eradicate the graphene-located cancer cells.

Thus far, this section of the book has focused on curing disease via nanomachines of some kind that circulate throughout the bloodstream, detect and maneuver to diseased cells, and terminate them efficiently. These solutions, however, only deal with illnesses that are caused by faulty individual cells, such as cancer or other similar diseases. Yet there are other disorders that stem from other sources. Many of these disorders, such as Parkinson's disease, originate in the brain. Chapter 14 [14] details how Prof. Parker's team has been able to electronically model four major diseases that stem from brain or other nerve-based abnormalities: Parkinson's disease, OCD, schizophrenia, and MS. Prof. Parker uses a collection of neuromorphic circuits consisting of both traditional CMOS circuitry and specific nanotechnologies to model each specific condition, allowing scientists to tweak and change the models in an attempt to create more effective treatments for these diseases. Like inorganic nanorobots, there are a number of problems still to be overcome, but the potential of these types of circuits to eventually assist in developing treatments for both these diseases and other neurological conditions is simply extraordinary.

Previous chapters of this book have already discussed how nanorobots can be used to treat disease within the body. What we have not yet addressed is how these swarms of nanorobots connect to the outside world. Chapter 15 [15] discusses linking medical nanorobots to pervasive computing. Given their intended environment, medical nanorobots must be very small and require minimal power. These requirements make it difficult to give them much onboard processing power or data storage. Instead, we can use wireless communication between each nanorobot inside the body and a central computer outside. This computer could handle much of the advanced algorithms and diagnostics required for the nanorobots to be successful in their mission. Outside computers could also guide nanorobots throughout their journey, relaying instructions to the nanorobots based on their present condition and location. This would reduce the amount of data storage required within each nanorobot for program data and processing. Doctors and scientists could also monitor data relayed from the nanorobots in real time in order to improve their own diagnostic abilities. These communication systems function in a very similar way as the Wireless systems discussed in the first part of this book. Connecting nanocomputing and Wireless computing systems will be a vital part of the future of nanomedicine, and one that should not be overlooked in the race to produce functioning nanorobot units themselves.

For most of its history as a field of study, nanomedicine has existed beneath a shadow of expectations. Nanomedicine—and nanotechnology in general—was supposed to cure disease, reverse the aging process, and usher in a new era of technological breakthroughs. As the authors of Chapter 16 [16] phrased it, nanomedicine was expected to bring about "radical transformations" in the way we both practice medicine and live our lives in general. In fact, sections of this book make similar claims for the coming future. Unfortunately, this has yet to occur. However, this is not necessarily an indictment of nanomedicine. Instead, the promises made in anticipation of sudden, revolutionary nanomedical innovation should not be taken literally. Instead, we should recognize their true purpose—a method to attract both scientists and funding to the cause—an effort that has been extraordinarily successful. In addition, nanomedicine is, in fact, slowly changing how we both research and practice medicine. Many of our initial predictions regarding the future of nanomedicine may still come to pass, as the field continues to work alongside other disciplines of medical research in an effort to improve the current state of medicine as a whole.

21.3 ETHICAL, PRIVACY, AND INTELLECTUAL PROPERTY ISSUES OF NANOMEDICINE AND WIRELESS COMPUTING

The medical industry is very old and well established. While they have had breakthroughs and new technologies in the past, nothing has changed the way that medicine is practiced to the degree that nanomedicine and Wireless computing likely will. For decades, the medical industry has followed the same code of ethics and adhered to the same standards of privacy. But with the advent of both Wireless and nanomedical technology, the old standards are no longer comprehensive. Chapters 18, 19 and 20 [18–20] discuss a plethora of ethical issues that accompany these technological breakthroughs. One such issue is that with the advances in wireless computing, we now have the ability to place nanoscale sensors throughout patients' bodies and collect private health data. In the interest of global health, should we share all collected data in one central location for anyone to analyze in the hopes of finding useful patterns or information within the mountains of data that they would then have to work with? Or should we keep patients' private data private, granting access only to each patient's personal doctors? On an even lower level, how much of a patient's data should we be allowed to record at all? Should there be a limit? Should patients even have to consent to noninvasive data collection, or should all collectable data be automatically harvested and shared to improve worldwide diagnostic effectiveness? There is no easy answer to these questions, nor to countless others like them. As the author says in Chapter 17 [17], in today's secular society we have no defined moral code we must follow. Society is always changing, as are our moral and ethical beliefs. We must do our best to look ahead and dictate policy, standards, and regulations as opposed to reacting to changes as they occur. At the same time, we cannot only rely on lawmakers or doctors to make these decisions for all of society. Doctors, patients, medical care and insurance providers, lawmakers,

and medical technology designers and manufacturers must all come together to develop a new code of ethics that will account for these changes in technique and technology.

21.4 CONCLUSIONS

This chapter briefly reviews the main topics discussed in the book. We discussed the importance of Wireless and nanoscale computing on the medical industry. First, we discussed how everything from cancer drug delivery and brain disorders to Markov decisions and cognitive disabilities are directly influenced by wireless computing and nanomedicine. Second, we investigated the microscopic world of nanomedicine. It also focuses on curing disease via nanomachines that circulate throughout the bloodstream, detect and maneuver to diseased cells, and terminate them efficiently. Last, we examine the ethical, privacy, and legal ramifications of nanomedicine and Wireless computing. We explain the existing laws and identify the area that we can improve in order to pave way for a sustainable technological development in the future. In summary, Wireless computing in nanomedicine is a very promising field with many challenges and issues that need to be addressed over the coming decades to make it truly viable.

ACKNOWLEDGMENTS

The authors acknowledge Pujal Trivedi, Arushi Gupta, and Alekhya Sai Nuduru Pati for devoting their time and effort toward this chapter.

REFERENCES

[1] A. Bhargava, M.M. Eshaghian-Wilner, A. Gupta, A.S. Nuduru Pati, P. Trivedi, “Chapter 1: Introduction,” Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications, by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[2] MM Eshaghian Wilner, Arushi Gupta, Shiva Navab, Alekhya Sai Nuduru Pati, Gaurav Sarkar,“Chapter 2: Nanocomputing and Cloud Computing,” “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[3] Linda Powers, Kui Ren, Janet Meiling Wang-Roveda, “Chapter 3: Pervasive Computing in Hospitals, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[4] Xiaojun Xang, “Chapter 4: Diagnostic Improvements treatment and care,” “Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications,” by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[5] Luis A. Castro, Edgar Chavez, Jesus Favela, Jessica Beltran-Marquez, Rene Navarro, Moises Perez, Marcela Rodriguez, Eduardo Quintana, "Chapter 5: Collaborative Opportunistic Sensing of Human Behavior with mobile phones," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[6] Franceli N Cibrian, Lizbeth Escobedo, Jose Mercado, Monica Tentori, "Chapter 6: Pervasive Computing to support people with cognitive disabilities," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[7] Zhuochen Ge, MM Eshaghian Wilner, Renjun Liu, "Chapter 7:Wireless Power for Implantable Devices A Technical Review," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[8] Daphney Stavroula Zois, Sangwon Lee, Murali Annavaram, Urbashi Mitra, "Chapter 8: Energy Efficient Physical Activity Detection in Wireless Body Area Networks," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[9] Shuping Liu, Anand Panangadan, Ashit Talukder, Cauligi S. Raghavendra, "Chapter 9: Markov Decision process for adaptive control of distributed Body sensor network," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[10] Amber Bhargava, MM Eshaghian Wilner, Wan Lee, Mike Schlesinger, Abbishek Uppall, Janet Cheung, "Chapter 10: An Introduction to Nanomedicine," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[11] MM Eshaghian Wilner, Sakhrat Khizroev, Gaurav Sarkar, Rakesh Guduru, Umang Sharma, "Chapter 11: NanoMedicine using Magneto-Electric Nanoparticles," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[12] Ido Bachelet, Noam Mamet, "Chapter 12: DNA Computation in Medicine," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[13] Farid Menaa, Sandeep Kumar Vashist, Adnane Adbelghani, Bouzid Menaa, "Chapter 13: Graphene Based Nanosystem for detection of Proteonic Biomarkers," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[14] Saeid Barzegarjalali, Rebecca Lee, Alice Parker, Sukanya Patil, "Chapter 14: Modeling brain disorders in silicon and nanotechnologies," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[15] Sylvain Martel, "Chapter 15: Linking Medical Robots to Pervasive computing," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[16] José J. López, Mathieu Noury, "Chapter 16: Nanomedicine 's Transversality -Some implications of the nanomedical paradigm," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[17] William Sims Bainbridge, "Chapter 17: Ethical Challenges of Ubiquitous Healthcare," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[18] Gaymon Bennett, Clark A. Miller, J. Benjamin Hurlbut, Heather M. Ross, "Chapter 18: The Ethics of Ubiquitous Computing in Healthcare," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[19] Fredrick Japhet Mtenzi, "Chapter 19: Privacy Protection of EHR in e-Healthcare Systems," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

[20] Ayush Chaudhary, Pujal Trivedi, MM Eshaghian Wilner, Arushi Gupta, Ben Shiroma, "Chapter 20: Ethical, Privacy, and Intellectual Property Issues in Nanomedicine," "Wireless Computing in Medicine, From Nano to Cloud with Ethical and Legal Implications," by Mary Mehrnoosh Eshaghian-Wilner, Hoboken, NJ: John Wiley & Sons, Inc., 2015.

INDEX

(Note: Entries in *italics* refer to figures, entries in **bold** refer to tables, n refer to foot notes)

Wireless Computing in Medicine: From Nano to Cloud with Ethical and Legal Implications, First Edition. Edited by Mary Mehrnoosh Eshaghian-Wilner.