Law, Governance and Technology Series 29

Brent Daniel Mittelstadt
Luciano Floridi *Editors*

The Ethics of Biomedical Big Data

Springer

Law, Governance and Technology Series

Volume 29

The *Law-Governance and Technology Series* is intended to attract manuscripts arising from an interdisciplinary approach in law, artificial intelligence and information technologies. The idea is to bridge the gap between research in IT law and IT-applications for lawyers developing a unifying techno-legal perspective. The series will welcome proposals that have a fairly specific focus on problems or projects that will lead to innovative research charting the course for new interdisciplinary developments in law, legal theory, and law and society research as well as in computer technologies, artificial intelligence and cognitive sciences. In broad strokes, manuscripts for this series may be mainly located in the fields of the Internet law (data protection, intellectual property, Internet rights, etc.), Computational models of the legal contents and legal reasoning, Legal Information Retrieval, Electronic Data Discovery, Collaborative Tools (e.g. Online Dispute Resolution platforms), Metadata and XML Technologies (for Semantic Web Services), Technologies in Courtrooms and Judicial Offices (E-Court), Technologies for Governments and Administrations (E-Government), Legal Multimedia, and Legal Electronic Institutions (Multi-Agent Systems and Artificial Societies).

More information about this series at http://www.springer.com/series/8808

Brent Daniel Mittelstadt • Luciano Floridi
Editors

The Ethics of Biomedical Big Data

Springer

Editors
Brent Daniel Mittelstadt
Oxford Internet Institute
University of Oxford
Oxford, UK

Luciano Floridi
Oxford Internet Institute
University of Oxford
Oxford, UK

ISSN 2352-1902 ISSN 2352-1910 (electronic)
Law, Governance and Technology Series
ISBN 978-3-319-33523-0 ISBN 978-3-319-33525-4 (eBook)
DOI 10.1007/978-3-319-33525-4

Library of Congress Control Number: 2016948203

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

Contributors

Margunn Aanestad is a Professor at the Department of Informatics, University of Oslo. She studied medical electronics engineering (combined B. Eng and M. Eng) at the University of Stavanger and received her Ph.D. on informatics from the University of Oslo. During the past decade, she has studied how healthcare institutions organize their information processes and how these processes impact service provision. Her research has a special focus on technologies related to inter-organizational, networked collaboration. She is a member of the Association of Information Systems. She has been a member of the editorial board of the *Scandinavian Journal of Information Systems* (2010–2013), *Information Technology and People* (since 2004), *Journal of the Association of Information Systems* (since 2014), and *Information and Organization* (since 2015).

Eric I. Benchimol is an Assistant Professor in the Department of Pediatrics and the School of Epidemiology, Public Health and Preventive Medicine at the University of Ottawa. He is also a pediatric gastroenterologist at the Children's Hospital of Eastern Ontario (CHEO) Inflammatory Bowel Disease Centre (cheo-ibd.ca, @CHEOIBD), a scientist at the CHEO Research Institute, and a scientist at the Institute for Clinical Evaluative Sciences (ICES). Dr. Benchimol conducts epidemiology, outcomes, and health services research using health administrative data. He is co-chair of the RECORD steering committee and helped develop the guidelines for the REporting of studies Conducted using Observational Routinely collected Data (RECORD). Dr. Benchimol is supported by a New Investigator Award from the Canadian Institutes of Health Research, Canadian Association of Gastroenterology, and Crohn's and Colitis Canada.

Paula Boddington has worked on diverse issues in applied ethics, focusing especially on ethical issues in clinical genetics and genomics, including problems concerning the sharing of personal medical information and scientific data. She has a particular interest in the intersection between questions in ethics with epistemology

and the philosophy of mind. Her degrees are in philosophy, psychology, and medical law. She has held posts at Bristol University, the Australian National University, Cardiff University, and the University of Oxford.

Markus Christen is a Senior Research Fellow at the Centre for Ethics of the University of Zurich and coordinator of the research network "Ethics of monitoring and surveillance". He is co-chair of the Human Brain Project's Ethics, Legal and Social Aspects Committee (ELSA). His research interests are in empirical ethics, neuroethics, ICT ethics, and data analysis methodologies. He has published almost 70 contributions in various fields (ethics, complexity science, and neuroscience) and he has authored or co-edited ten books.

Paul De Hert is a full-time Professor at the Vrije Universiteit Brussel (VUB), Associate Professor at Tilburg University, and Director of the Fundamental Rights and Constitutionalism Research Group (FRC) at VUB. After having written extensively on defence rights and the right to privacy, De Hert now writes on a broader range of topics including elderly rights, patient rights, and global criminal law.

Josep Domingo-Ferrer is a Distinguished Professor of Computer Science and an ICREA-Acadèmia Researcher at Universitat Rovira i Virgili, Tarragona, Spain, where he holds the UNESCO Chair in Data Privacy. His research interests include data privacy and data security. He holds Ph.D. and M.Sc. degrees in Computer Science from the Autonomous University of Barcelona; he also holds an M.Sc. in Mathematics. He has co-authored over 350 papers and five patents. He is a Fellow of IEEE and an Elected Member of Academia Europaea.

Bogdan Draganski is Consultant Neurologist at the University Hospital Lausanne, Director of the neuroimaging laboratory LREN, and Associate Professor at UNIL. He pioneered computational anatomy research by conceiving the speculative idea that local structure in the mature human brain may change in response to training and learning. His ongoing projects are in the field of neurodegenerative disorders with particular emphasis on the identification of surrogate imaging biomarkers in the presymptomatic phase of disease as an aid to the development of new therapeutic approaches.

Luciano Floridi is Professor of Philosophy and Ethics of Information at the University of Oxford, where he is the Director of Research and Senior Research Fellow of the Oxford Internet Institute, Governing Body Fellow of St Cross College, Distinguished Research Fellow of the Uehiro Centre for Practical Ethics, Faculty of Philosophy, and Research Associate and Fellow in Information Policy of the Department of Computer Science.

Urs Gasser is Professor of Practice at Harvard Law School and Executive Director of the Berkman Center for Internet and Society at Harvard University. His research and teaching activities focus on interdisciplinary information law, policy, and

society issues, with a current emphasis on comparative privacy in the age of Big Data and the Internet of Things. He has authored numerous articles and books, including *Interop: The Promise and Perils of Highly Interconnected Systems* (with John Palfrey). Dr. Gasser is also a Guest Professor at KEIO University (Japan) and was Visiting Professor at the University of St. Gallen (Switzerland). He has received several awards for his work at the intersection of law, technology, and markets.

Bryce Goodman is a graduate of the University of Oxford, Deep Springs College and Singularity University, and is currently pursuing a graduate degree in Philosophy and Data Science at the Oxford Internet Institute under the supervision of Luciano Floridi. His research is at the intersection of technology, philosophy, and innovation. A serial entrepreneur, his honors include Harvard Business School's "Best New Venture" (2011), Forbes' "30 Under 30: Energy & Industry" (2014), and World Economic Forum's "Technology Pioneer" (2015).

Dara Hallinan studied law in the UK and in Germany and completed a Master in Human Rights and Democracy in Italy and Estonia. Since May 2011, he has been a researcher at Fraunhofer ISI in Karlsruhe. The focus of his work is the interaction between new technologies – particularly ICT and biotechnologies – and society. He is writing his Ph.D. under the supervision of Paul De Hert at the Vrije Universiteit Brussel on the possibilities presented by data protection law for the better regulation of biobanks and genomic research in Europe.

Edward Hockings is a campaigner and researcher. He has held positions with Big Brother Watch and Action for Children and has a B.A. in Philosophy (Sussex) and an M.A. in Ethics and Law (Kings College London). He was the first person to obtain evidence of the 100,000 Genome Project and campaigns for higher levels of transparency and public engagement in the biosciences and information governance with EthicsandGenetics.org, of which he is the Founding Director. His work has been covered by the BBC News, The Guardian, The Independent, The Observer, and The Times.

Klaus Hoeyer's background is in social anthropology and medical ethics. His research interests include regulatory science, ethics as policy work and the social organization of biobanks and transplant services. He has published in a variety of journals and is the author of "Exchanging Human Bodily Material: Rethinking Bodies and Markets" (Springer).

Linda F. Hogle is a Professor of Medical Social Sciences at the University of Wisconsin-Madison. Her research deals with emerging medical technologies including regenerative medicine, precision medicine, and biomedical devices. Her work deals with themes of how novel technologies come to be standardized (or not) and more recently, changing forms of evidence in data-driven biomedicine.

Nelson Hosley is a graduate student in Philosophy at Brandeis University. He received his M.Sc. in Philosophy of the Social Sciences from the London School

of Economics, where he served as the Journal Coordinator for the Rerum Causae: Journal of the LSE Philosophy Society (2013). Before that, Nelson studied philosophy and sociology at the University of Pittsburgh, where he was co-editor of the Pitt Sociology Journal (2010–2011).

Sinéad M. Langan is a Senior Lecturer at the London School of Hygiene and Tropical Medicine (LSHTM) and honorary consultant dermatologist at St John's Institute of Dermatology, London. She leads a large programme of work using electronic medical record and administrative data which aims to answer key questions relevant to understanding herpes zoster natural history and informing vaccination policy. She also uses the power of routine data sources to provide answers for important research questions related to a range of skin diseases. She is co-chair of the RECORD steering committee and has co-led the development of guidelines for the REporting of studies Conducted using Observational Routinely collected Data (RECORD). Dr. Langan is supported by a Clinician Scientist award from the National Institute for Health Research.

Sandra Liede (b. 1977, LL.M. and Ph.D. Candidate, University of Helsinki) is a lawyer specialized in biomedical law and works as Senior Officer, Legal Affairs of Biobanking, at the National Supervisory Authority for Welfare and Health, Finland. Her research interests focus on the commercial factors influencing biomedical research and science and health policy solutions. She is a legal expert in a Finnish government-led working group, which has just recently published a national genome strategy for Finland.

Peter Mills' work has consistently explored the intersections of biomedical science, ethics, and public policy. He is currently Assistant Director at the Nuffield Council on Bioethics, an independent UK organisation that examines and reports on ethical issues relating to developments in biological and medical research. From 2007 to 2010, Peter was Head of Human Genetics and Bioethics at the UK Department of Health. As well as heading the secretariat for the Human Genetics Commission, the UK Government's independent advisory body on the implications of developments in human genetics, Peter has also represented the UK government at the UNESCO Intergovernmental Bioethics Committee (IGBC) and the Council of Europe Bioethics Committee (DH-BIO). Before moving to the Department of Health, Peter led a number of high-profile policy initiatives at the Human Fertilisation and Embryology Authority, concentrating on ethical, legal, and psychosocial aspects of developments in assisted conception and human embryo research. Some time before that, Peter read Philosophy, Politics, and Economics at Trinity College, Oxford, and went on to receive a Ph.D. in Philosophy from the University of Warwick.

Brent Daniel Mittelstadt is a Postdoctoral Research Fellow at the Oxford Internet Institute, University of Oxford. Since 2014, he has held a Junior Research Fellowship with St. Cross College. His current work examines the ethics of learning

algorithms as used in personal data analytics. Prior to this, he worked on the 'Ethics of Biomedical Big Data' project with Prof. Luciani Floridi to map the ethical landscape surrounding mining and sharing of biomedical and health-related 'Big Data' across research and commercial institutions. He has also conducted ethical foresight of emerging medical information and communication technologies, including personal health monitoring devices and 'smart' environments designed to support dementia care and 'ageing at home'. His research falls broadly within the philosophy and ethics of information, computer ethics, and medical ethics.

Stuart G. Nicholls is a Clinical Investigator and Methodologist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute, and Research Associate at the School of Epidemiology, Public Health and Preventive Medicine at the University of Ottawa. Having trained in both the basic and social sciences, his research sits at the intersection of ethics, social science, health policy, and health services research. At CHEO, Dr. Nicholls works to support and facilitate researchers using health administrative data, clinical data repositories, and research datasets in pursuit of the objectives of the Ontario Child Health SPOR Support Unit.

Annika Richterich is an Assistant Professor in Digital Culture at Maastricht University's Faculty of Arts and Social Sciences (NL). Her latest research focuses on digital materialism as well as services based on search engine data; currently, she conducts field research on innovation and learning practices in Dutch hacking communities. From a methodological perspective, she is interested in qualitative, empirical media research, while she has likewise critically engaged with debates concerning big data and Digital Humanities.

Christoph Schickhardt is postdoctoral researcher in biomedical ethics and coordinator of the project "DASYMED: Big data in Systems Medicine" at the National Center for Tumor Diseases at the University Hospital of Heidelberg (Germany). From 2013 to 2014, he coordinated the interdisciplinary consortium "Ethical and Legal Aspects of Whole Genome Sequencing" (EURAT). Christoph studied philosophy at the Universities of Pavia, Italy, and Lausanne, Switzerland, and was awarded a Ph.D. degree in Ethics by the University of Düsseldorf (Germany) in 2011. He teaches philosophy at the universities of Heidelberg and Bamberg (Germany).

Espen Skorve is a Postdoctoral Fellow at the Department of Informatics, University of Oslo. He studied informatics, sociology, and pedagogics (B.Sc. and M.Sc.) at the University of Oslo, and received his Ph.D. here as well. His research interests are related to the complexity of large-scale knowledge and information-infrastructures, with a focus on diversity in knowledge practices and how this diversity is reflected in design and development implementation and use of information technologies. Prior to joining academia he worked in IT-consulting and operations, primarily within the finance business.

Tade Spranger is Associate Professor at the Faculty of Law and head of the Junior Research Group "Norm-Setting in the Modern Life Sciences" of the Institute of Science and Ethics (IWE), University of Bonn, Germany. He is member of the Senate Commission on Genetic Research of the German Research Foundation (DFG). He has published more than 270 publications on National and International Life Sciences or Technology Law, Intellectual Property Law, and German Administrative and Constitutional Law.

Rochelle E. Tractenberg is a tenured Associate Professor at Georgetown University. Her primary appointment is in the Department of Neurology, and she has secondary appointments in the Departments of Biostatistics, Bioinformatics & Biomathematics and Rehabilitation Medicine. A professional biostatistician since 1997, she earned a Ph.D. in Cognitive Sciences/Psychology from the University of California, Irvine (1997), a M.P.H. emphasizing Biostatistics and Biometry from the California State University at San Diego (2002), and a Ph.D. in Measurement, Statistics, and Evaluation from the University of Maryland, College Park (2009). Her biomedical research interests are in measurement and outcomes in challenging biomedical contexts (e.g., estimating change in "cognitive function"; testing measurement invariance for complex neuropsychological constructs) and clinical trial design that features these challenging outcomes. She is also an active scholar of teaching and learning, focusing on cognitive theoretic contributions to learning in graduate and postgraduate education, and instruction in statistics and research ethics in particular. She is the Vice-Chair of the Committee on Professional Practice of the American Statistical Association.

Aaro Tupasela (b. 1972, DSocSc 2008) is a sociologist specialized in STS and works as an Associate Professor of ethical, legal, and social aspects of biobanking at the University of Copenhagen. He is a board member and former chair of the European Sociological Association's Sociology of Science and Technology Network, and also served as a member and chair of the Nordic Committee on Bioethics.

Polyxeni Vassilakopoulou is a Postdoctoral Fellow at the Department of Informatics, University of Oslo. She studied industrial engineering (combined B. Eng and M. Eng) at the Technical University of Crete, and operations research at Columbia University (obtained an M.Sc. as a Fulbright Scholar). She received her Ph.D. from the National Technical University of Athens. Her research interests are related to information systems for complex work settings with a dual focus on system's design and systems' appropriation and use. Empirically, her research work is focused in healthcare. Prior to joining academia, she worked in management consulting for over a decade successfully leading large-scale projects of information technology enabled interventions within the services sector (financial services, public sector, and social services). She is a chartered engineer and member of the Association of Information Systems.

Effy Vayena is a Professor of Health Policy at the University of Zurich, where she leads the Health Ethics and Policy Lab. From 2000 to 2007, she was a technical officer at the World Health Organization (WHO), working on ethical and policy issues relating to health research ethics, reproductive health ethics. She is a consultant to WHO on several projects, and visiting faculty at the Harvard Center for Bioethics, Harvard Medical School. In 2015–2016, she is a Fellow at the Berkman Center for Internet and Society at Harvard Law School. Her current research focus is on ethical and policy questions in personalized medicine and digital health. At the intersection of multiple fields, she relies on normative analyses and empirical methods to explore how values such as freedom of choice, participation, and privacy are affected by recent developments in personalised medicine and in digital health. She is particularly interested in the issues of ethical oversight of research uses of big data, ethical uses of big data for global health, as well as the ethics of citizen science. She has published widely in major journals in medicine, public health, health policy, and ethics.

Henrik Walter is Professor of Psychiatry, Psychiatric Neuroscience, and Neurophilosophy and Director of the Research Division of Mind and Brain at the Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Germany. He is chair of the Ethical Advisory Board of the Human Brain Project. His clinical-oriented research focuses on system neuroscience in psychiatry, in particular with respect to schizophrenia and depression using methods of cognitive neuroscience, neuroimaging, and genetics. He is also working on the cognitive neuroscience of volition, emotion regulation, and social cognition and in the field of neuroethics, neurolaw, and philosophy of psychiatry.

Eva C. Winkler is senior physician in oncology and head of the program "Ethics and Patient-oriented Care in Oncology" at the National Center for Tumor Diseases at the University Hospital of Heidelberg, Germany. She is also the speaker of the interdisciplinary consortium "Ethical and Legal Aspects of Whole Genome Sequencing" (EURAT). Prof. Winkler is a board-certified internist working in oncology in-and outpatient care for 14 years, attending of the Department of Medical Oncology and is heading the Clinical Cancer Program for Neuroendocrine Tumors. She holds a Ph.D. in cancer research from the University of Heidelberg as well as a Ph.D. in Medical and Healthcare Ethics from the University of Basel, Switzerland.

Simon Woods is Senior Lecturer and Co-Director of the Policy Ethics and Life sciences Research Centre at Newcastle University. Woods has a longstanding interest in medical ethics and law and in the ethics and regulation of research; he has been involved in the ethical review of research through national and international committees. His research explores the social and ethical aspects of new and emerging biotechnologies and has been a co-investigator in several EU projects with a focus on rare disease genomics.

J. Patrick Woolley is a Postdoctoral Fellow at the University of Oxford. After doing research in genomics and proteomics for several years, Patrick came to Oxford for his Ph.D. Interested in the changing relationship between science, ethics, and society, his dissertation examined post- and neo-Kantian influences in Albert Einstein's writings on ethics and religion. His postdoctoral studies on metaethics focus on the importance of consent in Rawls' Kantian constructivism and social contract theory. His work with HeLEX examines conditions of consent in current governance of biomedical data.

Introduction

Brent Daniel Mittelstadt and Luciano Floridi

1 Background

Modern information societies are characterised by mass production of data about humans. Digital technologies, including online services and emerging ubiquitous computing devices, can track behaviour to a greater degree than ever possible (Markowetz et al. 2014). Referred to as 'Big Data', this scientific, social and technological trend has helped create destabilising amounts of information, which can challenge accepted social and ethical norms. As is often the case with the cutting edge of scientific and technological progress, understanding of the ethical implications of Big Data lags behind.

Practices centred on the mass curation and processing of personal data can quickly gain a negative connotation which, in a way similar to what has happened in the public debate over genetically modified organisms (cf. Devos et al. 2008), places potentially beneficial applications at risk through association with problematic applications. A 'whiplash effect' can occur, by which overly restrictive measures (especially legislation and policies) are proposed in reaction to perceived harms, which overreact in order to re-establish the primacy of threatened values, such as privacy. Such a situation may be occurring at present as reflected in the debate on the proposed European Data Protection Regulation currently under consideration by the European Parliament (Wellcome Trust 2014), which may drastically restrict information-based medical research utilising aggregated datasets to uphold ethical ideals of data protection and informed consent.

Ethical foresight may reduce the probability of 'regulatory whiplash' by informing public debate through improved understanding of the moral potential of emerging technological applications and data practices. Analysis is required of

B.D. Mittelstadt (✉) • L. Floridi
Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford OX1 3JS, UK
e-mail: brent.mittelstadt@oii.ox.ac.uk; luciano.floridi@oii.ox.ac.uk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_1

issues and concepts known to be relevant (Mittelstadt and Floridi 2016), including informed consent, research ethics, privacy, confidentiality, anonymity, data ownership and digital divides. Issues of social justice, social profiling, collective rights, trust between data subjects and processors, intellectual property and access rights may also prove relevant through foresight.

To contribute to this process, this book presents cutting edge research on the new challenges of biomedical Big Data technologies and practices. The entries contained in this volume assess the transformative effects of Big Data on ethical norms and accepted practice. The volume offers an overview of the ethical problems posed by aggregation and re-purposing of biomedical datasets around issues such as privacy, consent, ownership, power relationships and digital divides. It discusses different approaches and methods that can be used to address these problems, particularly through policy and regulation. The book contains 19 original contributions on the analysis of the ethical, social and related policy implications of the analysis and curation of biomedical 'Big Data', written by leading experts in the areas of biomedical and technology ethics, Big Data, privacy, data protection, profiling and information ethics. The book advances our understanding of the ethical conundrums posed by biomedical Big Data datasets and analytics, and shows how policy-makers can address these issues going forward.

2 Big Data

Broadly, Big Data can refer to (1) the *process* of analysing 'big' data sets, and (2) the *datasets* themselves. 'Big' can be defined variably in terms of quantities of electronic size (gigabytes, terabytes, petabytes, etc.), entries, individuals or events represented by the data, or alternatively in relation to the techniques and technologies currently available for analysis. The latter approach defines 'big' in procedural rather than quantitative terms, by connecting the size of the dataset to its complexity, understood in terms of the computational or human effort necessary for analysis (e.g. Costa 2014; Dereli et al. 2014; Fan and Bifet 2013; McNeely and Hahm 2014; National Science Foundation 2014; Terry 2012, p. 389). In other words, the data are 'Big' because they are difficult to sort and analyse with existing computing technologies.

While helpful for bridging the space between analysis processes and datasets, this approach suggests data that is 'Big' now may not be so in a year or a decade due to advances in computing technology and analysis procedures (Floridi 2012; Liyanage et al. 2014, p. 27). Although not semantically problematic (as adjectives describing technology tend to be relative, e.g. fast internet 10 years ago is slow internet today), this nevertheless poses a technological solution to an epistemological query by making the definition of 'Big Data' relative in relation to technical and analytical capacities. 'Big Data' becomes data that is difficult to analyse due to its size and complexity. This also suggests that more or better computing will enable us to

'get ahead' of the data and analyse all of it meaningfully again, as we did prior to the current era of Big Data. However, the exponential growth of data (Bail 2014, p. 465) suggests this is unlikely to occur, a point that further reinforces the view that Big Data describes a break with prior practice. Explicit consideration of historical context reduces the fluidity of the definition; in other words, labelling a study as 'Big Data' recognises the technical and analytical barriers faced at the time it occurred. Such fixed labelling may be important in ex-post ethical analysis.

Recognising these implications of a purely technical definition, it may be helpful to consider also the perceived value of Big Data as suggested in the types of analysis it allows. Boyd and Crawford (2012, p. 663) suggest Big Data is valuable due to the "capacity to search, aggregate, and cross-reference large data sets." Similarly, according to Floridi (2012), a unique feature of Big Data is the possibility of identifying small patterns and connections in quantitatively large (and often aggregated) datasets. 'Small patterns' refer to connections between entries within the dataset, meaning connections are found within a subset of entries in a much larger dataset.

3 Biomedical Big Data

In biomedical research, the analysis of Big Data has become a major driver of innovation and success. Epidemiology, infectious diseases, and genomics and genetics (Heitmueller et al. 2014; Kaye et al. 2012), are already deeply affected (Floridi 2012). 'Biomedical Big Data' refers to the emerging technologically-driven phenomena focusing on analysis of aggregated datasets to improve medical knowledge and clinical care. This area has gained significant attention due to a combination of two factors. On the one hand, there is the huge potential to advance the diagnosis, treatment, and prevention of diseases as well as foster healthy habits and practices (Costa 2014). On the other hand, there is the obvious, inherent sensitivity of health-related data and the implicit vulnerability and needs of those potentially requiring treatments (Pellegrino and Thomasma. 1993). Academically and commercially valuable biomedical big data can exist in many forms, including aggregated clinical trials (Costa 2014), genetic and microbiomic sequencing data (Mathaiyan et al. 2013; McGuire et al. 2008; The NIH HMP Working Group et al. 2009), biological specimens, electronic health records and administrative hospital data. Such data can be held in biobanks, cyberbanks and virtual research repositories (Costa 2014, p. 436; Currie 2013; Majumder 2005, p. 32). Compared with traditional forms of storage, such repositories tend to assemble aggregated datasets explicitly for research purposes with "virtually unlimited opportunities for data linkage and data-mining" (Prainsack and Buyx 2013, p. 73) due to the sheer scale of the datasets (Steinsbekk et al. 2013, p. 151).

Data can also be generated explicitly or covertly via social media applications and health platforms (Costa 2014; Lupton 2014, p. 858), emerging 'personal health

monitoring' technologies (Mittelstadt et al. 2011, 2013) including wearable devices (Boye 2012), home sensors (Niemeijer et al. 2010) and smart phone applications, and online forums and search queries. The latter, for example, enable public health and outbreak tracking (Butler 2013; Costa 2014, p. 435). Other data come from 'data brokers' which collect, process, store and sell intelligence based on a variety of medical and health-related data sourced from social media, online purchases, insurance claims, medical devices and clinical data provided by public health agencies and pharmacies, among others (Terry 2012, 2014).

Analysis of these data types can be undertaken for numerous purposes, including development of clinically useful predictive models (Choudhury et al. 2014, p. 3), longitudinal and cross-sectional effectiveness and interaction studies of pharmaceuticals (Tene and Jules Polonetsky 2013, p. 246), and long-term 'personal health monitoring' (Boye 2012; Mittelstadt et al. 2014; Niemeijer et al. 2010). Broadly, these data may foster understanding of health disorders and the efficiency and effectiveness of treatments and health systems and organisations. They also create repositories for public health and information-based research (Safran et al. 2006, p. 2; Steinsbekk et al. 2013, p. 151). With that said, clinical applications are not guaranteed (Lewis et al. 2012). While promising on many fronts, biomedical Big Data, and the findings derived from it, may raise a host of ethical concerns stemming from the sensitivity of data being manipulated and the seemingly limitless potential uses and repurposing, and implications of data that concern individuals as well as groups. Precisely these concerns are the motivation for this volume that contributes new perspectives on key ethical challenges raised by Big Data methods in biomedical research.

4 Structure of the Volume

In the following pages these and related issues concerning philosophy, ethics, governance and policy are explored in much greater detail over 14 chapters representing the cutting edge of research on the ethics of biomedical Big Data. The book is divided into six parts. Part I addresses how Big Data creates imbalances between individual and collective interests, in particular through the re-purposing of non-medical data for medical purposes, which must be corrected. Part II continues this theme by examining imbalances specifically related to privacy interests and the shortcomings of data protection law in the context of a particular type of biomedical Big Data: large sample genomics research. Part III examines the imbalance between individual protection via informed consent and the social benefits of research created by Big Data processes that fundamentally challenge the feasibility of single-instance consent. Part IV explores how issues such as those raised in the first half of the volume concern the governance of biomedical Big Data repositories. Part V examines complementary requirements to governance structures surrounding challenges to professional norms, codes of conduct and the need for new ethical duties among researchers in response to Big Data methods of research. Part VI

then concludes with broader overviews of the ethics of biomedical Big Data, which serve as guidance for foresight analysis of new Big Data methods, platforms and processing contexts.

4.1 Part I: Balancing Individual and Collective Interests

Medical research fundamentally operates on a balance of individual and collective goods; the research participant willingly grants access to her body or records for the sake of advancing medical knowledge and, thereby, social good. The research participant willingly accepts risks to her body, well-being, or privacy for the sake of others. Much of biomedical Big Data involves re-use and re-purposing of existing clinical records, trials data, biobank samples and non-medical behavioural data. Re-purposing creates new risks for the individuals and groups described by the data or affected by the outcomes of the resulting research. Four entries to the volume describe challenges arising from re-use of data and the balance between individual and collective interests in biomedical Big Data.

Effy Vayena and Urs Gasser unpack the need for a new ethics framework to address the unresolved challenges of the intersection of traditional biomedical data and non-biomedical data. Data from Google searches, social media content, loyalty card points and similar applications can have high biomedical value. Insights can be drawn into a person's current health, future health, attitudes towards vaccination, disease outbreaks within a country and epidemic trajectories in other continents despite the data not explicitly describing health parameters. Their contribution highlights the 'digital phenotype' project to demonstrate a Big Data ecosystem in action, before unpacking the key components, design requirements and normative elements of a 'data ecosystem' ethics framework that responds to the challenges arising for re-purposing of non-biomedical data.

Annika Richterich expresses similar concerns around the need for ethical reflection on the use of non-biomedical data for epidemiological surveillance (or 'infodemiology'). Her contribution critiques methodological developments in epidemiological surveillance of influenza via data from internet sources. She describes the history of epidemiological surveillance from the 1980s, noting that influenza surveillance has traditionally relied on strictly biomedical data, typically from clinical and virological diagnosis or mortality rate statistics. Google Flu Trends is examined as a case study to examine the ethical implications of entanglements between public health services, emerging digital technologies and corporate objectives in internet-based epidemiological surveillance.

Klaus Hoeyer moves from epidemiological surveillance to epidemiological research facilitated by the ease of linking health and demographic data in Denmark. He notes that Denmark is often portrayed as an 'epidemiologist's dream' due to the ease of linking medical and non-medical datasets covering the country's entire population, without needing to obtain consent. Rich datasets are created by a health service with a remit to gather more data, of better quality, on more

people ('intensified data sourcing'). Discussion of the ethics of such 'intensified data sourcing' unfortunately tends to focus on the rights of the individual in terms of privacy and autonomy, despite data collection taking place at population level. He concludes that new modes of ethical reasoning and policy are required that originate in an understanding of actual data practices, which necessitate attention for the interests of the population as a whole.

To conclude Part I, Edward Hockings expands considerations of individual and public goods beyond concerns with re-purposing of data for public health projects through a critical analysis of policy developments in information governance and the biosciences. He examines the shift from rights-based approach to the adjudication of competing claims that is implicit in the justification of many biomedical Big Data research projects that create or re-use large clinical datasets. Five initiatives (the Clinical Research Practice Datalink, the Health and Social Care Information Centre, the 100,000 Genome Project, the introduction of personalised medicine, and the relaxation of the information governance regulatory regime) are considered that demonstrate how individual interests to privacy and confidentiality are not treated as inviolate rights, but rather goods to be balanced with societal goods, such as benefits to the economy or medical knowledge. This balancing act is shown to have demonstrable impact on current policy governing biomedical Big Data projects. An approach to policy and governance along deliberative and democratic lines is advocated in response to the novel ethical challenges of placing greater emphasis on economic benefits of biomedical research.

4.2 Part II: Privacy and Data Protection

Continuing with the policy focus on which Part I ended, Part II examines issues of privacy and data protection legislation applied to a particular type of biomedical Big Data: large sample genomics research. Medical data are traditionally held to be a particularly sensitive type of personal data, necessitating stricter limitations on its processing by third parties. However, as argued by Dara Hallinan and Paul de Hert, conceiving of biomedical Big Data repositories as strictly data repositories is misleading in the case of genomics research. Many biobanks contain biological samples and specimens alongside data derived from their sequencing or testing. Current European data protection law draws a distinction between samples and data: biological specimens are not seen to consist of or contain data, although data derived from their manipulation is considered personal data. Hallinan and de Hert argue against this conception, insisting instead that samples do in fact contain personal data. They argue that the forthcoming General Data Protection Regulation must be adapted to better protect the interests of donors to biobanks, in particular concerning genomics research. Specifically, biological samples must be seen to contain data in the form of DNA.

Hallinan and de Hert's contribution implicitly concerns appropriate boundaries for genetic privacy as enacted through data protection law. Bryce Goodman offers a

related perspective. His contribution explicitly examines shortcomings in the right to genetic privacy, which can prove a barrier to large-scale genomic research. His examination leads to both a negative and positive claim about the value of genetic privacy. Negatively, he asserts that genetic privacy is not intrinsically valuable, and that the barriers to genomic research posed by an unqualified right to genetic privacy are not justified. Positively, he concludes that genetic research is supported by the principle of respect for autonomy contained within the right to genetic privacy.

4.3 Part III: Consent

As suggested in discussions of the right to genetic privacy, individual interests and rights can prove both a barrier and enabler to biomedical Big Data. Nowhere is this more accurate than in the context of informed consent, a hallmark of medical research ethics. The two contributions to Part III describe the challenges and potential solutions faced in adapting informed consent for biomedical Big Data repositories and research studies.

The adaptation of models and mechanisms of informed consent to biomedical Big Data research has not proven easy. Traditionally, consent is case or jurisdiction specific; individuals agree to undergo a particular procedure or participate in a particular study following in-depth consideration of its merits and risks, assisted by informed medical professionals. As noted by J. Patrick Woolley, this single-instance model does not translate well to Big Data research defined by data re-use, aggregation and linking of medical and non-medical datasets. A gap has opened as a result in which policymakers have failed to create standard methods to address the ethical, legal and social issues (ELSI) arising in the Big Data environment. In his chapter, Woolley presents a view of governance where dataflow itself, not institutional or national boundaries, is taken as the *de facto* framework for research, and where metadata on consent play a central role in how data are governed. Types of consent are identified as an ideal starting point for the development of ELSI metadata procedures that assure data production, dissemination, and reuse stay within the boundaries of participants' and researchers' expectations.

Markus Christen, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, and Henrik Walter see similar problems with single instance consent, which they believe to be conceptually incompatible with exploratory Big Data research in which all possible hypotheses to be tested are not known at the time consent is obtained. They propose 'open' or 'broad' consent as an alternative when restrained by a clear framework defining legitimate and illegitimate types of research for a particular dataset or sample. The Human Brain Project is discussed as an example to show the difficulty of defining such a framework for Big Data research. A framework is currently being developed within the Project for access to multitude of clinical data related to brain diseases based on the conviction that many neurological and psychiatric disorders and diseases are ill-defined in terms of underlying mechanisms.

The inherent uncertainty of this type of research gives rise to ethically relevant consequences that must be considered when designing new consent mechanisms for biomedical Big Data.

4.4 Part IV: Ethical Governance

Biomedical Big Data often involves biobanks and repositories of medical data. As consent and data protection mechanisms adapt to new opportunities for data re-use, the fiduciary relationship between data subjects and repositories becomes critical. Ethics and governance committees increasingly manage access to biomedical Big Data resources. In deciding who is given access to the data, and in what format, governance bodies are trusted to protect and balance the interests of individual data subjects, the scientific community, commercial actors and the general public. Doing so requires consideration of the range of issues identified across this volume. The four entries in Part II address challenges of ethical governance of biomedical Big Data resources.

Picking up where Part I left off, Simon Woods applies Prainsack and Buyx's (2013) framework of 'solidarity' to two cases studies of research into rare diseases, which often requires combining genetic sequencing with medical records and natural history data. Solidarity emphasises the public good of data sharing and research in discussions around governance and consent. Woods argues that solidarity can provide the basis for governance of biomedical Big Data, although in some cases the model presumes too much good will on the part of data subjects. A need for a more collaborative approach to governance is called for in rare disease research to give research participants an opportunity to be able to negotiate the conditions of participation in research.

Polyxeni Vassilakopoulou, Espen Skorve, and Margunn Aanestad continue the focus on genetic biomedical Big Data with an examination of emerging tensions related to data ownership and sharing in global genetic data repositories hosted by both public and private institutions. They describe the on-going controversies around collecting and sharing genetic mutation data on the *BRCA1* and *BRCA2* genes: the creation of the Breast Information Core (BIC) database in 1995, the decision by Myriad Genetics to stop sharing information in 2004, the subsequent reaction from the community through the "Sharing Clinical Reports Project" and "Free the Data" initiatives and the recent creation of the open ClinVar repository and the public-private BRCA Share resource. Multiple rationalities guiding positions on data ownership and sharing are identified. Their contribution turns to prior work in collective actions and governance of the commons to as a way to find common ground on questions related to equity, efficiency and sustainability. Answering these questions is critical to the design of context appropriate governance for genetics repositories.

In her contribution, Paula Boddington analyses the ethics of managing public accessibility and private control of biomedical Big Data from the perspective of

theories of communication. A comparison is drawn to ethical issues arising from the communication of personal and familial medical and genetic information. She argues that the situated and personal communication of knowledge generates ethical considerations which may clash with impersonal or system-driven understandings of data, and the related ethical responsibilities experienced by individuals. Nissenbaum's theory of privacy as contextual integrity is used to assess the importance of channels of communication in determining responsibilities within data management and governance, including how channels of dissemination can contribute to or assuage feelings of disempowerment among data subjects.

Aaro Tupasela and Sandra Liede conclude the part with a critical examination of the right of access to data held in biobanks granted by Finland's Biobank Act (688/2012). Biobanking and data sharing infrastructures pose new ethical and legal dilemmas in the interpretations of data subject rights in relation to processing of personal data. The Act requires biobanks to provide, upon request, information regarding data which may have clinical (actionable) relevant for the data subject's personal health. Such concerns are common to biomedical Big Data repositories holding identifiable (personal) data. While a right to access may combat feelings of disempowerment among data subjects, governance mechanism do not currently exists in Finland through which common access standards and practices could be implemented. The management of data, research results and incidental findings in biobanks is becoming, however, an increasingly significant challenge for all biobanks and the countries which are in the process of drafting policy and regulatory frameworks for the management and governance of big data, public health genomics and personalised medicine. Tupasela and Liede's examination of the Finnish case speaks to the challenges faced across Europe and elsewhere in terms of how to govern and coordinate the management of biomedical Big Data.

4.5 Part V: Professionalism and Ethical Duties

Ethical governance is, however, not sufficient by itself to guarantee ethically responsible research. Adaptations to the professional responsibilities of researchers and medical practitioners involved in the collection, aggregation, linking and analysis of medical data are also necessitated by the emergence of Big Data research. The three contributions in Part V detail some of the adaptations required, in particular concerning practices required to promote transparency.

Christoph Schickhardt, Nelson Hosley and Eva C. Winkler build an analytical and ethical framework to assess the theoretical and practical feasibility of an ethical duty for researchers to share pre-publication data with the scientific community. They ask whether researchers have a prima facie duty to share pre-publication data and, if so, which constraints and interests must be considered to determine the force of the duty in particular contexts. Data sharing is seen as a requirement to fulfil the role of science as a social good advanced through promotion and adoption of scientific knowledge. The authors analyse the concept of data sharing and clarify

what data sharing might imply in practice. Their framework calls for context-specific assessment of stakeholder interests. It is argued that these interests, which often conflict with the prima facie duty to share data, are determined in part by the normative-informational environment in which data producing researchers (to whom the prima facie duty to share data applies) are usually situated.

Stuart G Nicholls, Sinéad M. Langan and Eric I. Benchimol are similarly concerned with the transparency in reporting of studies using large-scale health-related datasets. Reporting of methods used in Big Data research studies are seen as practically beneficial as it allows for appropriate peer review and critical evaluation of studies; facilitates reproduction and replication of research findings; may help to reduce waste, and avoid redundancy and unnecessary repetition; and may facilitate public trust in scientific research. In parallel with the previous chapter, transparent reporting is seen as an essential component of researcher integrity. The chapter reports on recommendations from the RECORD Statement (REporting of studies Conducted using Observational Routinely-collected Data) on transparent reporting in studies using routinely collected health data.

Rochelle Tractenberg addresses the related topic of guidelines for professional practice concerning research ethics, which can conceivably include a duty to share collected data. She argues that professionals in computing and statistics typically do not receive training in responsible conduct of research, which creates a professionalism gap due to the important role of these professionals in biomedical Big Data. The emergence of biomedical Big Data as a cross-disciplinary phenomenon means the sort of professional norms or codes of conduct typically associated with individual professions will not necessarily emerge. Tractenberg examines the state of professional guidelines in the United States. She argues that dominant Federally-funded training programmes in 'responsible conduct of research' are unlikely to support the development of appropriate professional norms for biomedical Big Data. An alternative approach is described that can support ongoing reflection on professional obligations and ELSI concerns, including those that have not yet been identified (a key focus given the uncertainty of hypotheses in exploratory Big Data research). Guidelines for professional practice from three statistical associations (American Statistical Association; Royal Statistics Society; International Statistics Institute) and the Association of Computing Machinery provide the basis of the approach advocated.

4.6 Part VI: Foresight

The volume concludes with three selections that provide a broader view of the ethical challenges faced in biomedical Big Data. Each builds upon current knowledge to identify critical points and themes foresight analysis of the ethics of specific Big Data methods, platforms and processing contexts.

Linda F. Hogle's contribution explores the ethical and political aspects of infrastructures that support Big Data projects for medical research and clinical care. She observes a reordering of relationships between patients, clinical and family caregivers, researchers and payers, with potentially long-term implications for concepts of autonomy and expertise, among others. As suggested in Klaus Hoeyer's contribution on 'intensive data sourcing' in Denmark, the transformations to medical relationships brought about by Big Data practices broadly represent a distortion of the traditional distinction between research and clinical care. Imagined futures of healthcare as 'personalised medicine' represent such a reordering of relations, wherein iteration is encouraged between data-driven clinical care and the underlying analysis of large, streaming datasets of routine medical data.

Pete Mills' contribution takes as a starting point the findings the Nuffield Council on Bioethics 2015 report 'The collection, linking and use of data in biomedical research and health care: ethical issues'. A key recommendation made in the report was for Big Data initiatives to negotiate context-specific moral requirements for data use with data subjects at a local level. The chapter unpacks how this recommendation can operate in practice, arguing that organising data initiatives as social practices that respect certain principles can help to establish and meet morally reasonable expectations about data use, by grounding them in a dynamic relationship between social norms, individual freedoms and professional duties.

In the volume's final chapter, Brent Daniel Mittelstadt and Luciano Floridi provide a systematic overview of the ethical concepts and issues relevant to Big Data analytics in general, and biomedical Big Data in particular. A thematic narrative is offered to guide ethicists, data scientists, regulators and other stakeholders through what is already known or hypothesised about the ethical risks of this emerging and innovative phenomenon. Five key areas of concern are identified: (1) informed consent, (2) privacy (including anonymization and data protection), (3) ownership, (4) epistemology and objectivity, and (5) 'Big Data Divides' created between those who have or lack the necessary resources to analyse increasingly large datasets. Critical gaps in the treatment of these themes are identified with suggestions for future research. Six additional areas of concern are then suggested which, although related have not yet attracted extensive debate in the existing literature. It is argued that they will require much closer scrutiny in the immediate future: (6) the dangers of ignoring group-level ethical harms; (7) the importance of epistemology in assessing the ethics of Big Data; (8) the changing nature of fiduciary relationships that become increasingly data saturated; (9) the need to distinguish between 'academic' and 'commercial' Big Data practices in terms of potential harm to data subjects; (10) future problems with ownership of intellectual property generated from analysis of aggregated datasets; and (11) the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources. Considered together, these 11 themes provide a critical foresight framework to guide ethical assessment and governance of emerging Big Data practices.

References

Bail, Christopher A. 2014. The cultural environment: Measuring culture with Big Data. *Theory and Society* 43(3–4): 465–482. doi:10.1007/s11186-014-9216-5.

boyd, danah, and Kate Crawford. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society* 15(5): 662–79. doi:10.1080/1369118X.2012.678878.

Boye, Niels. 2012. Co-production of health enabled by next generation personal health systems. *Studies in Health Technology and Informatics* 177: 52–58.

Butler, Declan. 2013. When Google got flu wrong. *Nature* 494(7436): 155–156. doi:10.1038/494155a.

Choudhury, Suparna, Jennifer R. Fishman, Michelle L. McGowan, and Eric T. Juengst. 2014. Big Data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience* 8: 239. doi:10.3389/fnhum.2014.00239.

Costa, Fabricio F. 2014. Big Data in biomedicine. *Drug Discovery Today* 19(4): 433–440. doi:10.1016/j.drudis.2013.10.012.

Currie, J. 2013. 'Big Data' versus 'Big Brother': On the appropriate use of large-scale data collections in pediatrics. *Pediatrics* 131(Suppl): S127–S132. doi:10.1542/peds.2013-0252c.

Dereli, Turkay, Yavuz Coskun, Eugene Kolker, Oner Guner, Mehmet Agirbasli, and Vural Ozdemir. 2014. Big Data and ethics review for health systems research in LMICs: Understanding risk, uncertainty and ignorance-and catching the black swans? *American Journal of Bioethics* 14(2): 48–50. doi:10.1080/15265161.2013.868955.

Devos, Yann, Pieter Maeseele, Dirk Reheul, Linda Van Speybroeck, and Danny De Waele. 2008. Ethics in the societal debate on genetically modified organisms: A (Re)quest for sense and sensibility. *Journal of Agricultural and Environmental Ethics* 21(1): 29–61. doi:10.1007/s10806-007-9057-6.

Fan, Wei, and Albert Bifet. 2013. Mining Big Data: Current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter* 14(2): 1–5.

Floridi, Luciano. 2012. Big Data and their epistemological challenge. *Philosophy & Technology* 25(4): 435–437. doi:10.1007/s13347-012-0093-4.

Heitmueller, A., S. Henderson, W. Warburton, A. Elmagarmid, A. Pentland, and A. Darzi. 2014. Developing public policy to advance the use of Big Data in health care. *Health Affairs* 33(9): 1523–1530. doi:10.1377/hlthaff.2014.0771.

Kaye, Jane, Liam Curren, Nick Anderson, Kelly Edwards, Stephanie M. Fullerton, Nadja Kanellopoulou, David Lund, et al. 2012. From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews Genetics* 13(5): 371–376. doi:10.1038/nrg3218.

Lewis, Cecil M., Alexandra Obregón-Tito, Raul Y. Tito, Morris W. Foster, and Paul G. Spicer. 2012. The Human Microbiome Project: Lessons from human genomics. *Trends in Microbiology* 20(1): 1–4. doi:10.1016/j.tim.2011.10.004.

Liyanage, H., S. de Lusignan, S.-T. Liaw, C.E. Kuziemsky, F. Mold, P. Krause, D. Fleming, and S. Jones. 2014. Big Data usage patterns in the health care domain: A use case driven approach applied to the assessment of vaccination benefits and risks. Contribution of the IMIA Primary Healthcare Working Group. *Yearbook of Medical Informatics* 9(1): 27–35. doi:10.15265/IY-2014-0016.

Lupton, Deborah. 2014. The commodification of patient opinion: The digital patient experience economy in the age of Big Data. *Sociology of Health & Illness* 36(6): 856–869. doi:10.1111/1467-9566.12109.

Majumder, M.A. 2005. Cyberbanks and other virtual research repositories. *Journal of Law, Medicine & Ethics* 33(1): 31–39. doi:10.1111/j.1748-720X.2005.tb00208.x.

Markowetz, Alexander, Konrad Błaszkiewicz, Christian Montag, Christina Switala, and Thomas E. Schlaepfer. 2014. Psycho-informatics: Big Data shaping modern psychometrics. *Med Hypotheses* 82(4): 405–411. doi:10.1016/j.mehy.2013.11.030.

Mathaiyan, Jayanthi, Adithan Chandrasekaran, and Sanish Davis. 2013. Ethics of genomic research. *Perspectives in Clinical Research* 4(1): 100. doi:10.4103/2229-3485.106405.

McGuire, Amy L., James Colgrove, Simon N. Whitney, Christina M. Diaz, Daniel Bustillos, and James Versalovic. 2008. Ethical, legal, and social considerations in conducting the human microbiome project. *Genome Research* 18(12): 1861–1864. doi:10.1101/gr.081653.108.

McNeely, Connie L., and Jong-on Hahm. 2014. The Big (Data) bang: Policy, prospects, and challenges. *Review of Policy Research* 31(4): 304–310. doi:10.1111/ropr.12082.

Mittelstadt, Brent Daniel, N. Ben Fairweather, Neil McBride, and Mark Shaw. 2011. Ethical issues of personal health monitoring: A literature review. In *ETHICOMP 2011 conference proceedings*, 313–321. UK: Sheffield.

Mittelstadt, Brent Daniel, N. Ben Fairweather, Neil McBride, and Mark Shaw. 2013. Privacy, risk and personal health monitoring. In *ETHICOMP 2013 conference proceedings*, 340–351. Denmark: Kolding.

Mittelstadt, Brent Daniel, N. Ben Fairweather, Mark Shaw, and Neil McBride. 2014. The ethical implications of personal health monitoring. *International Journal of Technoethics* 5(2): 37–60.

Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. The ethics of Big Data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics* 22: 303–341. doi:10.1007/s11948-015-9652-2.

National Science Foundation. 2014. Critical techniques and technologies for advancing Big Data Science & Engineer (BIGDATA) – Program Soliciation NSF 14-543. http://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf

Niemeijer, A.R., B.J. Frederiks, I.I. Riphagen, J. Legemaate, J.A. Eefsting, and C.M. Hertogh. 2010. Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: An overview of the literature. *International Psychogeriatrics* 22: 1129–1142.

Pellegrino, Edmund D., and David C. Thomasma. 1993. *The virtues in medical practice*. New York: Oxford University Press.

Prainsack, Barbara, and Alena Buyx. 2013. A solidarity-based approach to the governance of research biobanks. *Medical Law Review* 21(1): 71–91. doi:10.1093/medlaw/fws040.

Safran, C., M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, D. E. Detmer, and With input from the expert panel (see Appendix A). 2006. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association* 14(1): 1–9. doi:10.1197/jamia.M2273.

Steinsbekk, Kristin Solum, Lars Øystein Ursin, John-Arne Skolbekken, and Berge Solberg. 2013. We're not in it for the money—Lay people's moral intuitions on commercial use of 'their' biobank. *Medicine, Health Care and Philosophy* 16(2): 151–162. doi:10.1007/s11019-011-9353-9.

Tene, Omer, and Jules Polonetsky. 2013. Big Data for all: Privacy and user control in the age of analytics. http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20

Terry, Nicolas. 2012. Protecting patient privacy in the age of Big Data. *The UMKC Law Review* 81: 385.

Terry, Nicolas. 2014. Health privacy is difficult but not impossible in a post-HIPAA data-driven world. *Chest* 146(3): 835–840. doi:10.1378/chest.13-2909.

The NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J.A. Schloss, et al. 2009. The NIH Human Microbiome Project. *Genome Research* 19(12): 2317–2323. doi:10.1101/gr.096651.109.

Wellcome Trust. 2014. Impact of the draft European Data Protection Regulation and Proposed Amendments from the Rapporteur of the LIBE Committee on Scientific Research. Wellcome Trust. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf

Part I
Balancing Individual and Collective Interests

"Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine"

Effy Vayena and Urs Gasser

Abstract In today's ever evolving data ecosystem it is evident that data generated for a wide range of purposes unrelated to biomedicine possess tremendous potential value for biomedical research. Analyses of our Google searches, social media content, loyalty card points and the like are used to draw a fairly accurate picture of our health, our future health, our attitudes towards vaccination, disease outbreaks within a county and epidemic trajectories in other continents. These data sets are different from traditional biomedical data, if a biomedical purpose is the categorical variable. Yet the results their analyses yield are of serious biomedical relevance. This paper discusses important but unresolved challenges within typical biomedical data, and it explores examples of non-biomedical Big Data with high biomedical value, including the specific conundrums these engender, especially when we apply biomedical data concepts to them. It also highlights the "digital phenotype" project, illustrating the Big Data ecosystem in action and an approach believed as likely to yield biomedical and health knowledge. We argue that to address the challenges and make full use of the opportunities that Big Data offers to biomedicine, a new ethical framework taking a data ecosystem approach is urgently needed. We conclude by discussing key components, design requirements and substantive normative elements of such a framework.

1 The Chiaroscuro Portrait of Big Data

The "Big Data" phenomenon has undoubtedly captured the psyche of modern society. It's an alluring idea, elusive and almost inescapable. Allure surrounds spectacular expectations for what Big Data can deliver, but it is elusive because

E. Vayena (✉)
Health Ethics and Policy Lab, Institute of Epidemiology, Biostatistics and Prevention, University of Zurich, Hirschengraben 84, Zurich 8001, Switzerland
e-mail: effy.vayena@uzh.ch

U. Gasser
Berkman Center for Internet & Society, Harvard Law School, Harvard University, 23 Everett Street, 2nd Floor, Cambridge, MA 02138, USA
e-mail: ugasser@gmail.com

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_2

of our inability to define what exactly Big Data is. And all this remains inescapable because living in today's digitized world puts us right at the heart of it. We generate data and metadata in massive amounts. Our world has fetishized quantification (Feiler 2014), and most aspects of our lives are entangled in data we generate, capture and use. We live most of our lives online, and we do business online. We order our food, manage our financial assets, fall in love and have our diseases diagnosed online, and all of this activity is captured as data. Big Data is all about us and all aspects of our lives.

Given the lack of consensus on a definition, we tend to understand Big Data by describing the data's key characteristics: variety, velocity, veracity, and volume.[1] But these features are not the substantive reason for the enthusiasm they spark; rather, we get enthusiastic because we see data as a source we can exploit. The most commonly employed metaphor for Big Data is that of oil: Big Data as a natural resource, spewing forth from each of us as we live digitally, quantifiable and monetisable (Watson 2014). "Personal data is the oil that greases the Internet," Somini Segupta argued in a *New York Times* op-ed in 2012. "Each one of us sits on our own vast reserves. The data that we share every day—names, addresses, pictures, even our precise locations as measured by the geo-location sensor embedded in Internet-enabled smartphones—helps companies target advertising based not only on demographics but also on the personal opinions and desires we post online" (Sengupta 2012).

Fundamental to the concept of Big Data are the data analytics and data mining techniques deployed to distil meaning from the data themselves. These tools enable important inferences and identification of non-obvious patterns in human behaviour or other structures in organizations and networks. It is the analysis and mining of these huge amounts of data that make Big Data powerful. The growing volume of data is a rich source from which to tease out information relevant to an ever-expanding list of societal, technological, scientific, political and personal issues. And the important links are everywhere: data from our online purchases reveals our preferences, our opinions, and our health status. Our Facebook "likes" alone can accurately predict our sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender (Kosinski et al. 2013), and it is the power of accurate prediction, says Jonathan Shaw, that makes Big Data "a big deal" (Shaw 2014). In his recent book, the somewhat apocalyptically-titled "*Dataclysm: Who we are when we think no-one is looking*," Chris Rudder suggests that "practically as an accident, digital data can now show us how we fight, how we love, how we age, who we are, and how we're changing" (Rudder 2014).

The positive side of this view is the notion that we are finally in a position to understand ourselves and the many facets of our being. The many data points

[1] IBM. The Four V's of Big Data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data

we produce, recording the details of our lives, give rise to overarching patterns of actions, associations and behaviours. This is the *chiaro* aspect, or the brightness, of Big Data. As in the popular renaissance technique of *chiaroscuro*, the brightness dominates and brings the object into focus—not the details, but the illuminated object in its entirety.

Such highly penetrative power to reveal sought-after patterns and notions of who we are raises a number of wicked ethical questions about Big Data. The questions span a wide spectrum, and together they are the *scuro* aspect, or the darkness. We do not use *scuro* to imply nefariousness or negativity necessarily; rather, these complex queries are an essential part of the portrait, but still one in shadow. There are questions about how our autonomy, privacy and identity may be affected by Big Data. For example, how will our social norms that safeguard these values be sustained, or perhaps altered? Are existing regulatory schemes suitable for the ethical complexities of the Big Data challenge—indeed, are regulatory mechanisms the answer at all (Christie et al. 2015)? Can the real potential of Big Data be exploited while we are still unable to answer very fundamental questions about our moral interaction with it (Mayer-Schönberger and Cukier 2013)? The greater the volume of data available to us, and the more uses we put them to, the more urgent the need to explore this part of the Big Data phenomenon. All of these questions, and many others, raise a single bigger one: whether the brightness and glory of Big Data are partly an illusion created by keeping these other issues in the shadow.

The central idea we pursue here is that Big Data cannot easily be boxed into clearly demarcated, functional categories. Depending on how it is queried and combined with others, a given data set can traverse categories in complex and unpredictable ways. So it appears limiting to attempt to address ethical challenges as fundamental as autonomy, privacy and justice solely through context-specific approaches. Contexts matter, of course and determine the specific articulation of a given ethical question and its respective answer, but we argue that context-specific solutions should be embedded into a more comprehensive and coherent ethical framework for the Big Data ecosystem.

Biomedical Big Data is traditionally a category of data with clear contours, subject to strong regulatory oversight, and it is a case in point. Below, we discuss important but unresolved challenges within typical biomedical data; then we explore examples of non-biomedical Big Data with high biomedical value, and the specific conundrums these engender, especially when we apply biomedical data concepts to them. We proceed with the discussion of the "digital phenotype," an illustration of the Big Data ecosystem in action and an approach believed likely to yield improved biomedical and health knowledge.

In the last part of the paper, we articulate some key elements that an ethical framework should contain if it were to adopt the Big Data ecosystem approach. We do not aim to provide a complete framework here; rather we seek to suggest an approach that we believe is better suited to the special challenges of Big Data and that presents promise in terms of its ability to guide us through nuanced, systematic solutions. It is our contention that this approach might shed light on the shadowed parts of the Big Data landscape, so we can capture most of its potential.

2 Typical Big Biomedical Data

No aspect of life is untouched by digitization and data capture, and health and biomedicine are perhaps particularly susceptible. Examples are legion: clinical care data, laboratory data, genomic sequencing data, and data from various other fields of biology ending in -omics (Nuffield 2015). Not only do we now generate unprecedented amounts of this data, but we are also making significant progress in the bioinformatics and analytics that allow us to apply it to further our health and biomedical knowledge. We speak, therefore, of Biomedical Big Data. The National Institutes of Health define it as follows:

> Biomedical Big Data is more than just very large data or large numbers of data sources. Big Data refers to complexity, challenges, and new opportunities presented by the combined analysis of data. In biomedical research, these data sources include diverse, complex, disorganised, massive and multimodal data being generated by researchers, hospitals and mobile devices around the world"[2] (NIH).

This definition implies different categories of activities within Biomedical Big Data. Some use cases include

(a) Analysis of data of the same type within the same source, e.g., a large genomic data set at a certain institution;
(b) Analysis of data of the same kind that are not in the same data source, e.g., genomic data from different centres; and
(c) Analysis of combined data of different sorts, e.g., genomic data and medical records;

though many more examples could be added, as the space of possible applications and uses seems almost unlimited.

Genomics is a particularly useful example. A recent study by geneticists and computer scientists compared data generation in three different domains—astronomy, social media (YouTube and Twitter) and genomics—and generated a headline-catching article warning us to brace for the genomic data flood (Stephens et al. 2015). While astronomy is traditionally a data-intensive field and data production in social media is exploding, genomic data are expected to surpass all others at high speed. A human genome has 3 billion base pairs; the sequence of a single genome constitutes about 100 Gigabytes (GB) of data. Given the decreasing costs of genomic sequencing and the current emphasis on the potential of genomic data for clinical and research applications, it is estimated that by 2025 between 100 million and 1 billion human genomes will be sequenced (Hayden Check 2015).

A total size of 100 GB for just one data set gives a sense of the massive volume of the data. Multiplying 100 GB by the billion people expected to be sequenced puts our sense of big in perspective, as it pushes our metrics to exabytes (10^{18} bytes), if not even further. As for the "biomedical" part, relating to biology and or medicine,

[2]Data Science at NIH. 2015. What is Big Data? https://datascience.nih.gov

genomics again meet the criteria, as genomic data result from the analysis of human DNA, and therefore constitute a core element of an individual's biological profile. They carry potentially important information about ancestry, health and diseases. The data are personal in the sense that each of us has a unique sequence, even if one person's genetic variation from another person's is only 0.1 %. The real benefit in terms of our understanding of health and disease requires pooling genomic data from many individuals and dwelling on differences and similarities. Such data may currently reside in different hospitals, research centres, and countries. To exploit them successfully, ideally the data should be pooled, allowing for higher statistical power and giving different research groups the chance to query them.

Several initiatives are underway to collect massive genomic data and facilitate its sharing among institutions. Meanwhile, though, it has become increasingly clear that genomics is just one piece of the larger jigsaw of human health, with several other –omics— proteomics, metabolomics and microbiomics to name only three— also being crucially important. The new popular paradigm of "precision medicine" has promoted the idea of an–omics driven medicine. Although often seen as synonymous with genetic medicine, the renewed ideal of precision medicine aims to draw on the various –omics to deliver more precise diagnosis and treatment. But the even more enticing prospect offered by precision medicine is that of using these –omics to predict disease and, ultimately, to prevent it. In this sense personalized medicine is a Big Data project.

Precision medicine advocates contend that progress can only be made if – omics are pooled in large repositories and analysed by different research teams (Auffray and Hood 2012; Hood and Flores 2012; Hood and Aufray 2013). But for this to happen, individuals have to authorize access to their data set, or even participate directly in the making of the new medicine by collecting it themselves and making it available for research. This sort of participation is enhanced by an increasing range of digital and mobile devices. Health apps, point-of-care diagnostics, wearable tracking technologies are all shaping the digital future of medicine, which is becoming increasingly personalized (Ginsburg 2014). The clamour to jump on the bandwagon grows steadily, but whether and how we might do this responsibly raises serious ethical questions that go beyond matters of what is scientifically or technologically possible.

A recent illustrative example of the vulnerability of Biomedical Big Data activities in the absence of supporting social norms is the NHS's care.data project. The project, aimed at aggregating all NHS patient data in order to facilitate medical research, has come to a dramatic and possibly terminal standstill (at least in its originally proposed form) over widespread public concerns about the consent processes and the protection of individual privacy (Mitchell et al. 2014). It is doubtful that Biomedical Big Data initiatives will ever really deliver on their promise unless such ethical challenges are dealt with adequately, in a manner that inspires public confidence and trust (Nuffield 2010, 2015; Juengst et al. 2012).

Not surprisingly, national and international bodies focusing on the diverse technical aspects of Biomedical Big Data initiatives have repeatedly highlighted the importance of identifying and exploring their ethical dimensions, and have urged

the broader scientific and ethico-legal community to provide guidance. Despite expressed interest in these ethical considerations, we still lack adequate policies and societal consensus even for questions from the early days of genetic medicine and biobanking (Vayena et al. 2008; Widdows 2013). Illustrative examples include issues of informed consent for biobank samples, appropriate biobank governance schemes, and sample and data ownership, to name just a few. We are no closer to consensus, either. The latter question has been answered very differently in various jurisdictions, and the moral underpinnings of these various judicial decisions remain unclear (Angrist 2007). As the number of data initiatives grows steadily, and collaborative projects (including data linking projects) become more common, such unresolved questions generate confusion, and ultimately receive hasty and *ad hoc* responses that may not always meet ethical requirements.

In August 2014 the US National Institutes of Health (NIH) updated its genomic research guidelines, requiring researchers funded by the NIH to post genomic data online for other researchers to use. While this requirement was an update of an existing policy, a key development was a further explicit requirement to obtain consent from study participants to share their data with other researchers.[3] The shortfall, as commentators have discussed, was that the NIH provided no guidance on what type of consent is appropriate, what other information it should include, whether it should be renewed, and whether it can be revoked (Van Noorden 2014). To comply with the guidelines and obtain consent for data sharing, the appropriate type of consent in the relevant case *must* be specified.

It is also important to re-examine the weight attributed to consent, especially in such large linking projects, which present endless possibilities for research and repurposing of data. The problem with consent in Biomedical Big Data scenarios is multifaceted, as the Mittelstadt and Floridi meta-analysis of academic literature discussing ethical aspects of Big Data indicates (Mittelstadt and Floridi 2016). The fundamental "impossibility of certainty concerning future uses of data," which is inherent to Big Data, is in sharp contrast with the traditional notion of informed consent, which cannot be "informed" at the time of consent as far as future and often unrelated investigations based on shared, aggregated, and reused data are concerned. Second, attempts to "fix" or "sidestep" traditional single-instance consent mechanisms by re-consent, blanket consent, tiered consent, or alternative models may either be impractical and costly, or trigger significant ethical concerns and, depending on jurisdiction, serious legal issues. In addition, a growing body of literature demonstrates that traditional techniques for anonymizing or de-identifying data, in ways which would dispense of some legal consent-related requirements by avoiding the regulatory triggers of "personal data" or "personally identifiable information" are generally ineffective (Narayanan and Felten 2014; Ohm 2010; Sweeney 2000). This is particularly the case in Big Data research environments

[3]National Institutes of Health. NIH Genomic Data Sharing Policy. August 27 2014. (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html).

that typically utilize sets of data containing many pieces of information for each individual, making each record unique and potentially identifable (de Montjoye et al. 2015; Almishari et al. 2014; Narayanan and Shmatikov 2008). For example, in a recent case, the National Institutes of Health rescinded public access to a database of aggregated genetic information because it was possible to confirm, with high statistical confidence, whether an individual was part of a population in a study about a specific medical condition (Homer et al. 2008; Felch 2008).

3 Non-biomedical Big Data of Great Biomedical Value

The section above focused on what we typically understand as biomedical big data and some of the key ethical questions generated by their uses. In this section we turn our attention to data that cannot be classified as biomedical data, and therefore, they are not governed by the same rules that apply to typical biomedical data. Although these data sets are different from traditional biomedical data, yet the results their analyses yield are of serious biomedical relevance. Analyses of Google searches, Wikipedia searches, social media content, loyalty card points and the like are used to draw a fairly accurate picture of not only our current health but also of our future health, our attitudes towards vaccination, disease outbreaks within our country, and even epidemic trajectories across other continents. In our diverse, evolving data ecosystem it is clear that data generated for a wide range of purposes unrelated to biomedicine still provides rich information about health. What follows is a series of illustrative examples, along with the respective ethical challenges they pose.

3.1 Loyalty Cards Points

The story of the American Department store Target is a widely publicized and striking example of the elasticity of Big Data. In 2012, the *New York Times* published a cover story exposing how the Target loyalty card data of a teenage customer led the company's marketing analysts to predict that she was pregnant. On the basis of her purchase history, Target sent a series of advertising coupons tailored to pregnancy needs to her home; her father then complained about the coupons, only to eventually learn that she was indeed pregnant (Duhigg 2012). This story has been recited numerous times, becoming synonymous in some circles with the creepy face of data analytics (Schneier 2014). And it is undoubtedly unsettling to have intimate personal health information visible to unknown, untrusted others. It is not only questions of harm that raise concerns; after all, this information might never be communicated beyond the database and might never cause social harm to the individual in question. Rather, it is the basic fact that this personal health information has become available without the person in question being aware of it, or having any control over its availability, that constitutes a fundamental privacy

invasion. It is evident that signing an agreement for a department store loyalty card does not currently constitute informed consent to generate and use health or behavioural records.

If we could, however, temporarily overlook the creepiness and privacy concerns, the compelling takeaway of this story is the evidence of a real-world capability to derive such personal health information from a shopping list. A pregnancy is typically diagnosed through a urine test or a blood test, on the basis of hormonal levels in one's body, and these hormonal levels are explicitly biomedical data. The ontology of biomedical data is mostly constructed by their source (the human body) and content (cholesterol values, genetics sequence, etc.). While a shopping list can hardly count as biomedical data under current definitions, nonetheless it enables a fairly accurate prediction of a biological event.

3.2 Social Media

Users of social media of all kinds—Facebook, Twitter, PatientsLikeMe, Dailystrength, etc.—share health-related information with commercial services and, in turn, with friends and sometimes even the public (Fox 2011). The content of such posts can be mined for a variety of health specific issues (Mandeville et al. 2014; Vayena et al. 2015), including adverse reactions to drugs, defined by WHO as

> . . . any response to a drug which is noxious and unintended, and which occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease, or for the modifications of physiological function (World Health Organization 2002).

National drug regulators such as the US Food and Drug Administration (FDA) are mandated to collect and evaluate these adverse events, to understand the biomedical processes that underpin them, it is imperative to collect data on such reactions and assess their magnitude and severity. However, the way in which pharmacovigilance, as this activity is called, is practiced today has serious limitations. The main deficiencies of the current system are serious under-reporting, poor reporting, and time lag between evaluation of reporting and action (Levinson 2012). Limited or poor data on adverse events poses risks to individual patients and is costly for health care systems (Heger 2015).

Data on adverse events are typical biomedical data. Although their uses are mainly in the aggregate, they include symptoms, they are linked to medical conditions and they convey serious information about individuals. For example, several studies have successfully demonstrated that Twitter posts, or "tweets," can be used for pharmacovigilance. Freifeld and colleagues used 6.9 million twitter posts (approximately 400 million of these are generated per day at time of writing) containing references to a medical product, and successfully identified adverse events relating to 23 conditions (Freifeld et al. 2014). While much remains to be done to fine-tune this method of pharmacovigilance, even regulatory authorities are starting to show interest in utilizing this approach (Heger 2015).

But are Twitter users aware that their data is being used for pharmacovigilance? Are they in agreement? Does the mere fact that a tweet is publicly available allow for any kind of use? Does our traditional public/private dichotomy hold in the online world? What responsibilities do those who collect such data bear for those whose data are collected; for example, what if a serious adverse event is detected that requires medical attention? Is the collector morally responsible for suggesting that an individual seek such attention? Or even for providing it herself (Kahn et al. 2014)? Some of these issues have long been debated in the standard biomedical research domain, and the ethics codes and regulations of health research dictate how biomedical research should be conducted when it involves human participants, their samples and their data. But existing processes and systems that try to protect autonomy, anonymity and privacy do not address sufficiently the Big Data uses as they relate to biomedical research.

For example, consider the principle of "informed consent" that underpins so much of the biomedical research ethics paradigm. Typically, biomedical data are obtained with the prior informed consent of the person providing them, and this consent is expected to do a lot of the ethical work, particularly to protect autonomy—albeit nominally in many cases. Despite the fact it has been empirically shown that informed consent *doesn't* always protect autonomy in biomedical research, research ethics codes still lack nuance in relation to the consent requirement (Manson and O'Neill 2007). Typically the informed consent documents are lengthy and written in technical language; they do not take advantage of online technologies and visualizations techniques that can facilitate meaningful engagement with the content. Far and foremost, informed consent, as practiced, is a static solution to a dynamic issue. People change their minds overtime, over their life course, in response to life events etc. Others may not need to understand every detail of a research project provided that the project meets ethical requirements and is conducted in a trustworthy environment. Nuance is necessary in both the process of consent as well as in its normative substance (Koenig 2014).

Back to Twitter, posts are not biomedical data *per se*, and are obtained online subject to a very broad agreement from the average user. It is hard to imagine that anyone who signs up on Twitter can predict at the point of "consent" (i.e., registration) the future content of her tweets, how that content is going to be used, and by whom (Vayena et al. 2013). If you asked her at the point of signing up, it is fair to assume that "pharmacovigilance" would be an unlikely response. Users blithely "agree" to various terms of service or privacy policies; is it then appropriate to expect a biomedical project using Twitter data to satisfy the consent requirement?

A related but separate issue is the online distinction between private and public (Gleibs 2014; Zimmer 2010). While this line is brighter and clearer in the physical world, the same is not true online, for reasons that have been detailed in the literature. Studies of social media users have shown that being online and sharing information about oneself is not necessarily the same as having decided to go public (O'Brien et al. 2015). Users may still have expectations of privacy while being active online, and this includes expectations not to be tracked or to have personal data (beyond actual postings on a social medium) used and shared with other entities. In

the words of Justice Sotomayor on a case of personal data, "it may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties."[4]

3.3 Mobile Devices

The number of smart phones is soon expected to surpass the world's population. These phones are equipped with sensors and allow geolocation data to be captured, an abundance of such data already exists, and such data are used in biomedicine and public health (Anema et al. 2014). The 2014–2015 Ebola crisis illustrated how cell phone data can be used in an infectious disease pandemic or other public health emergency; mobile phone data in affected regions was exploited for contact tracing and other public health surveillance activities detecting human mobility (Wesolowski et al. 2014). The main obstacles for using such data, even in the context of an international health emergency, were the nature of the data and the fact that they were the property of telecommunication companies. Typically such data are not used for public health purposes, and, unsurprisingly, policies allowing swift release of the data are lacking (The Economist 2014). In certain circumstances, however, they are immensely valuable. While cell phone data are not biomedical in nature, they nonetheless allow identification of disease trajectories more accurately than a given set of biomedical data at that time, and in 2014–2015 they enabled health authorities to act efficiently and trace those at risk. In this case, the use and content of the data once again belong in different categorical boxes.

Another striking example from cell phone data use is a recent study demonstrating how mobile phone data such as daily locations and time spent on the phone could be used to identify the severity of depression symptoms. On the basis of these data researchers could detect depression symptoms much more accurately than they could using standard questionnaires (Saeb et al. 2015). One of the study authors framed the success of the study in an insightful, but problematic, way:

> The significance of this [finding] is we can detect if a person has depressive symptoms and the severity of those symptoms without asking them any questions. We now have an objective measure of behaviour related to depression. And we're detecting it passively. Phones can provide data unobtrusively and with no effort on the part of the user (Paul 2015).

Simply having a cell phone generates enough of a digital trail to allow detection of health symptoms, but no terms of service from a telecommunications company is currently likely to contain a clause referring to such potential uses. Cell phone

[4]U.S. v. Jones, 132 S.Ct. 945, 957 (2012) (Sotomayor, J., concurring).

contracts and patient consent serve different legal and moral purposes—and, indeed, it is unlikely that one could opt out from data collection of this sort. The only such option might be not to carry a mobile phone at all.

The example of depression is interesting, as undiagnosed depression is highly prevalent, and the example presents a possible scenario where cell phone data could be used to screen populations for depression, alert individuals of risk, or even introduce emergency measures if algorithms detected risk of suicide or self-harm. Such scenarios were spotlighted by a recent debate over an app produced by the Samaritans charity that could detect suicide risk from people's Twitter feeds; there was such a backlash over privacy and stigmatization concerns that eventually the app was withdrawn (Samaritans 2014).

4 The Digital Phenotype

Our increasing interaction with digital technologies and devices, and their effects on us and our behaviour, have given birth to a new concept: the digital phenotype. It is based on the idea of the extended phenotype, first introduced by Richard Dawkins, who argued that our phenotype cannot be limited strictly to our biological processes (Dawkins 1982). Instead, our interactions with the environment and the ways in which we modify it are part of our wider phenotype. Capturing and understanding these interactions allows greater understanding of how we function.

Jain and colleagues take this idea further, opening the notion of the phenotype to our daily digital interactions (Jain et al. 2015). We interact with personal digital technologies, we modify them and they affect us, and they constitute a major part of our environment; as such, they are natural extensions of our phenotypes. In constructing the digital phenotype, Jain *et al*. are after the data captured by such interactions and what they tell us for health and disease.

> Through social media, forums and online communities, wearable technologies and mobile devices, there is a growing body of health-related data that can shape our assessment of human illness. Such data have substantial value above and beyond the physical exam, laboratory values and clinical imaging data—our traditional approaches to characterizing a disease phenotype (Jain et al. 2015).

The data can reveal behaviours and patterns caused by biological processes, and by analysing them we can explore those processes.

In this understanding the digital phenotype serves as a normative guide that aims to offer:

(a) An alternative, but not exclusive, approach to the standard biomedical paradigm that typically starts with a fixed hypothesis about biological processes and aims to collect evidence to refute or approve it; and
(b) A more unified take on Big Data, whereby all data that can be captured about or from a person can contribute to understanding biology, health and disease.

The "digital phenotype" gives name to several calls over recent years to exploit all kinds of data in relation to the individual (Murdoch and Detsky 2013; Ayres et al. 2014). Several epistemic arguments have been offered in support, including the identification of new hypotheses, real time insight, assumption-free insight, etc. Weber et al. visualized these arguments in a graph entitled "the tapestry of potentially high value information sources that may be linked to an individual for use in health care" (Weber et al. 2014). The wefts in the tapestry are data: electronic health records, genomic sequence data, credit and loyalty card data, and Facebook and Twitter use data, among others. In this view, combined analysis of all data fills the gaps inevitably created when just a subset of the data—a small piece of the tapestry—is analysed.

Scientists advocating the idea of data-rich biomedical research have also identified the obstacles to such an approach. Beyond technical issues, most of which are likely to be overcome, are the more complex ethical and regulatory problems. The data sets that should be accessed to create these digital phenotypes are typically in silos, locked in systems that lack not only technical, but also legal and policy interoperability. They are controlled by a variety of entities with different governing policies, including varying terms of service, intellectual property arrangements and licensing schemes, privacy policies etc. Different proprietary interests and inhomogeneous privacy concerns are among the key drivers of such non-interoperable policies, and are commonly cited obstacles for the feasibility of the digital phenotype (Pentland et al. 2013).

Such obstacles are created partly by the nature of the possibilities unfolded by Big Data; but also, crucially, by our attempts to tame the Big Data phenomenon with ethical frameworks and regulatory approaches designed to address conceptually different research models and practices. This can have unfortunate outcomes: projects like the digital phenotype can continue evolving within constraints that make it impossible to realise their full potential. This will undermine trust in Big Data research, for fear it is approached for the wrong reasons. On the other hand, various entities invested in making Big Data deliver on its promise (and in getting returns on their financial investment) may find ways to proceed with biomedical projects without adequate ethical safeguards in place. These run the risk of triggering public outcry and harsh regulatory responses that can dramatically stymie biomedical research innovation. What we need, therefore, is a proactive, innovative, ethically robust response to the challenges and opportunities of big data in biomedicine.

5 Towards a New Ethical Framework

5.1 Vision for a New Framework

The discussion in the previous sections brings to light three key points: First, the Big Data phenomenon creates new challenges in the context of typical biomedical

data, which are not sufficiently addressed by the legal and ethical best practices that emerged in the age of (relatively) small data. Second, although biomedical data are categorized on the basis of their source and content, big data from non-biomedical sources can be used for biomedical purposes. The challenge is therefore shifted from the source or content of the data to its use, creating thorny ethical questions that test the effectiveness of traditional consent-based safeguards. Third, moving from risks to opportunities, the greatest benefit for biomedicine will come precisely from linking and exploiting all kinds of data relating to an individual—and, by extension, all such data from all populations. Taken together, these three developments fundamentally challenge standard biomedical ethics creating the need for a new approach with a new framework at its core.

Such a framework must be aimed at establishing common practices to govern the creation, collection, storage, processing, internal sharing, analysis, dissemination, and re-use of data in biomedical research, as well as long-term access to it. Spanning the full life cycle of data, the framework must confront head-on the taxonomical problem inherent in Big Data, placing enhanced ethical scrutiny on the actual uses of the data rather than on abstract data taxonomies. The need for this new biomedical research ethical framework is pressing. If it were to be developed and adopted it could also inform the next generation of legal requirements in key areas such as privacy protection, where traditional instruments such as notice and consent are equally challenged, and new consensus is yet to emerge on the design of appropriate safeguards for the new data environment.

A fully fleshed out ethical framework is beyond the scope of this article and requires not only further research and facilitation, but also a concerted effort by the various stakeholders involved (Gasser et al. 2015). Here, as a starting point for further discussion, we offer some thoughts on key components of the envisioned framework including design requirements and some substantive normative elements.

Going beyond (but certainly drawing on) existing ethics guidelines in biomedical research activities, the proposed new ethics framework would be composed of at least three analytically distinct, albeit interacting, components:

- A *set of ethical norms* at its core, guiding researchers, research ethics committees and other stakeholders engaging in biomedical research in a Big Data environment. Such norms would incorporate or build upon existing best practices (Secretary's Advisory Committee on Human Research Protections 2015), initially taking the form of guidelines and over time crystallizing into codes of practice. Central issues to be addressed—as outlined briefly below—include privacy, transparency, and accountability.
- *Process guidelines* for decision-makers engaged in biomedical projects. A process approach to ethical decision-making in Big Data research would emphasize the ethical responsibility of researchers (as well as others involved), while simultaneously providing support in operationalizing the ethical norms. Again, the framework would build upon existing and emerging guidelines (Markham and Buchanan 2012).

- *Tools and resources* providing practical support to ethical biomedical research in Big Data environments. The tools in the box would evolve over time: initial resources could include new methodologies and approaches to evaluate privacy risks when sharing research data (e.g., DataTags,[5] an automated tool for assessing privacy risks and generating custom data handling policies). These would be complemented by educational materials and resources for researchers and review boards, forming the basis of a knowledge foundation on ethics and Big Data research.

5.2 Design Requirements

We envision the following baseline characteristics of the new ethical framework.

First, the framework has to be rooted in an *ecosystem perspective*, overcoming the traditional boundaries and taxonomies that no longer apply, including some of the distinctions between different types of data that have lost their meaning in the age of Big Data. An ecosystem approach also ensures that all relevant actors—some of which are new entrants to the world of biomedical research—are part of the picture and share ethical responsibility as appropriate. This novel perspective is required in order conceptually to capture the emergence of projects such as the digital phenotype. It also indicates that the framework includes the full lifecycle of biomedical research data, including creation, storage, sharing, aggregation, and re-use.

A second requirement of the envisioned framework is *interoperability*. By default, it should be designed to be applicable to a broad range of technological and data infrastructures, both now and in the future, and to interact meaningfully with different organizational settings and diverse policy environments (Palfrey and Gasser 2012). To give an example at the organizational level, the proposed new framework should have interfaces with existing IRB/ethics review processes, but also interoperate with emerging bodies outside traditional research institutions, such as consumer review boards, participant-led research review boards (Vayena and Tasioulas 2013) or recently proposed personal data cooperatives (Hafen et al. 2014). At the policy layer, the new framework should facilitate the working together of both generic policies on ethical human research in the age of Big Data—ranging from guidelines to legal requirements—and context-specific norms, such as emerging standards for ethical research in social media environments (Rivers and Lewis 2014). The recent proposal to create a Safe Harbor Framework for International Ethics Equivalency is an example of how one might seek to increase both organizational and policy interoperability for biomedical research at the international level (Dove et al. 2014).

The third design requirement of the envisioned new ethics framework is its *ability to evolve over time*. As the earlier parts of this paper illustrate, we live in a technologically fluid environment. The technology that enables today's biomedical

[5] DataTags. 2015. The President and Fellows of Harvard College. http://datatags.org/

research was not available a few years ago, and it continues to evolve rapidly, as the example of next-generation predictive algorithms illustrates. A new ethics framework for biomedicine has to incorporate mechanisms of learning that respond to such technological advancements and to changing user behaviour (i.e., the ways in which we as individuals and groups interact in the digital space and create new data of potential biomedical value), as well as to evolving social norms and values around data, privacy, and research.

5.3 Substantive Key Elements

We focus here on three normative elements: (1) ethical use and privacy protection; (2) data governance; and (3) transparency and accountability.

5.3.1 (1) Ethical Use and Privacy

The ethical framework for Big Data use in biomedicine needs to be focused on and calibrated for *actual uses* of the data, rather than exclusively on its collection at the source or raw content. The blurring lines between the traditional categories of data (O'Neill 2013; Schwartz and Solove 2011) used for biomedical research and the challenges faced by conventional consent mechanisms are the main reasons for an emphasis on usage (Cate and Mayer-Schönberger 2013), which encompasses various stages in the data life cycle model. The ethics norms developed as part of the framework—and its supporting processes and tools—must improve on current approaches, which from a public interest perspective may put unnecessary burdens on activities such as public health research (for example, if consent was not obtained for a particular use); but which do not guarantee that personal privacy is indeed protected, as in the case of many private sector Big Data activities where terms of service do not refer explicitly to specific research uses (Kahn et al. 2014). With Biomedical Big Data projects such as the digital phenotype, different domains and differently regulated data sets have to find unifying operating principles.

Assessing the ethical use of data involves a risk/benefit assessment for any use. For instance, the proposed framework would provide or further develop tools and techniques such as a "privacy impact assessment" (PIA),[6] with the aim of mapping the entire spectrum of *privacy risks*. A number of emerging guidelines in the biomedical Big Data space propose PIA as a way of ensuring proportionate safeguards in data uses (Global Alliance for Genomics and Health 2015), but commentators have already suggested that the spectrum of risk is growing and evolving,

[6]PIA: A formal process which assists organizations in identifying and minimizing the privacy risks of new projects or policies that make use of Data. The assessment involves working with people within the organization, with partner organizations, and with the people affected to identify and reduce privacy risks.

putting additional pressure on standard PIA. For example, privacy challenges go beyond *tangible* (and measurable) harms such as material harm, to *intangible* harms such as exclusion or reputation damage, to more *abstract* ones such as social stratification or loss of trust (Polonetsky et al. 2014). A narrow focus on assessing tangible risk will only capture part of the risk spectrum. The new taxonomies of risk require privacy assessment to be redesigned and sensitive to these newly formed privacy risks that may be greatly damaging to individuals and to wider society.

In parallel, the assessment of benefits will also have to be adapted. A recent proposal for general Big Data uses introduced the concept of Data Benefit Analysis (FPF 2014). This focuses on assessing variables like the nature of the benefit, the identity of the beneficiary, and the likelihood that the benefit will be achieved. Ultimately, this more nuanced, structured analysis of potential benefits must be weighed against potential privacy harms. Biomedical uses of any data set would undergo such an assessment irrespective of the entity using the data.

Finally, the new ethical framework has to take into account changing and increasingly nuanced privacy attitudes and individual behaviours. Today, an individual's right to privacy is often portrayed as antagonistic to the health-related public goods that can result from increased openness. From this point of view, an individual's right to privacy may have to be routinely infringed in order to promote the public good of genomic research. There is a legitimate public interest in health knowledge, it is believed, that justifies such infringements in order to contribute to the good of health.

This picture of the relationship between the right to privacy and the public good of health is too crude. The legitimate public interest in research does not sideline the individual right to privacy (Vayena and Gasser 2016). In fact, it is often that very interest that allows research to take place. For example, people may be prepared to waive their privacy rights in order to secure personal and communal health benefits, and several studies have shown that people who care about their privacy (strongly believing that it is their right) are willing to sacrifice some of their privacy in order to participate in genomic research (Oliver et al. 2012). People negotiate their privacy in ways that reflect not only their views about privacy, but also the circumstances of their lives. These vary dramatically amongst people and over the life cycle.

5.3.2 (2) Data Governance

A second core set of issues that span the different components of the proposed ethics framework are those related to *data governance*.

Behind the seemingly technical question of data governance lies the more fundamental issue that the framework must tackle: the nature of the ethical, organizational and legal mechanisms providing opportunities for data subjects to be involved in the entirety of what happens with the data sets. This stretches beyond one's involvement in the use of one's own data to, for example, consideration of whether access to the entire data set should be granted for specific purposes; how

benefit-sharing or intellectual property is negotiated; and ultimately how one may want to use one's own data for one's own purposes.

One particular difficulty in this context is presented by those companies that enable users to collect personal health or other data, but which do not allow them to access their data sets or exercise even basic forms of control over their use. As the digital phenotype approach matures, individuals may need such access for their own personal health benefit; and for the larger biomedical project, access to such data will be needed at population level. In Switzerland, for example, the idea has been discussed of granting an individual Constitutional right to obtain an electronic copy of data concerning oneself (Gächter and Werder 2015). If such a right were granted, in principle individuals would be in a position to determine re-use of their data, including making them available for the digital phenotype project.

Evolving data portability requirements are pointing in a similar direction. Examples include those introduced in the EU's Draft Data Protection Regulation, which give individuals whose personal data are processed electronically and in a structured, commonly used format the right to obtain a copy of that data for further use.[7] One extension of this approach, which seeks to empower individuals, is a recent proposal to establish national personal data cooperatives owned and governed by citizens, independent from governments or corporations (Hafen et al. 2014).

In addition to incorporating next generation data governance norms that empower data subjects and—more ambitiously—potentially facilitate the creation of new norms, the envisioned framework would also provide organizational and tool-based support (such as, for instance, software-aided privacy risk assessment systems for researchers) to address data governance issues at the practical level. Researchers at the Berkman Center for Internet & Society at Harvard University, for example, have analysed a large number of data use agreements as part of an ongoing National Science Foundation project, Privacy Tools for Sharing Research Data[8]; from these, legal best practices can be distilled, potentially also informing technical solutions such as modular license generators.

5.3.3 (3) Transparency and Accountability

A third fundamental element of the ethics framework is a renewed focus on *transparency and accountability*. Big Data research in general, and data-driven research in the biomedical context in particular, is highly specialized and complex. Inherent information asymmetries between experts and researchers on the one hand and research subjects on the other are further amplified by the absence of legal obligations or robust practices aimed at information users, as seen in some of the privacy

[7] Article 18: Council of the European Union. 2015. Draft Data Protection Regulation. http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf

[8] Harvard School of Engineering and Applied Sciences. 2014. Privacy Tools for Sharing Research Data. http://privacytools.seas.harvard.edu/

sector examples provided above. There is growing consensus among experts that the solution to these structural information asymmetries cannot be resolved by standard disclosure practices, i.e., just by "throwing more information" at data subjects.

The envisioned framework would therefore have to move beyond good information practices to facilitating the development of supporting mechanisms—what we might call "tools for transparency." These could include dashboards, visualization techniques, and the like, similar to those that have been emerging in the consumer privacy space (Gasser 2015). Such mechanisms effectively inform individuals about uses of data that relate to them. This need for educational and translational information practices only increases as Big Data analytics advance and algorithms become self-learning and potentially even opaque to their creators.

Transparency norms and tools are fundamental dimensions—and often even prerequisites—of accountability. They can also serve as instruments for assessing behaviour and performance of all actors involved in a system. They therefore facilitate not only the approval of certain uses (e.g., through a system of independent review boards) but also the monitoring of those uses by institutions and individuals.

As the complexity of information flow increases, transparency by design will also enable automated accountability tools to be developed (for example, automated systems monitoring whether uses of data are compliant with relevant policies and user preferences). Scholars have been pointing in this direction of user-friendly accountability tools in biomedicine for some time, even before the Big Data era. O'Neill has observed that cumbersome accountability mechanisms even if intended to manifest trustworthiness they do not necessarily engender trust (O'Neill 2002).

While the discussion about research accountability in the age of Big Data is only now beginning, one could envision how the proposed new ethical framework would encourage and support the development of additional accountability mechanisms that supplement baseline legal liability as the ultimate accountability tool. Responsibility as pertaining to following either internally or externally established norms and practices, or responsiveness to other stakeholders as an outward-looking dimension of accountability, are other possible dimensions of a more advanced accountability approach that can inform the field of biomedical Big Data research, among others. Borrowing from other thematic contexts such as multi-stakeholder governance, the current emphasis on accountability as (often internal) compliance with codes of conducts, ethics guidelines, good practices, etc. could be supplemented by outward-facing accountability mechanisms such as responsiveness to stakeholders and the public at large, to name just one tool that is available in the governance toolbox.

6 Conclusion

Big Data comes with a big promise for society; this holds particularly true for biomedicine. However, a broad social consensus about what is desirable and permissible and where the limits of data-driven research should be set has yet to emerge. Unsurprisingly, given the lack of such normative guidance, we are currently

in the phase of patchwork approaches when regulating uses of Big Data in different contexts. Fragmented and ad hoc regulatory solutions are unlikely to allow Big Data to deliver on its promise in general or on its promise in biomedicine in particular. Moreover they also create gaps that leave privacy and identity vulnerable to violations, and societal trust in research at risk. This is the kind of double loss we must avoid. Developing a proactive, innovative and ethically robust response to the challenges and opportunities of Big Data is of paramount importance and urgently needed.

We have argued here that a new ethical framework should be developed for the use of Big Data in biomedicine. We sketched the basic features of such an interoperable framework, which would build upon existing and emerging guidelines to set forth norms, while also articulating practices and providing tools that help researchers, review boards, individuals, and the public at large to navigate the thorny ethical questions that we face today. Such a framework should take into account the larger seismic shifts in today's digitally connected ecosystem, including the blurring lines among traditional categories and taxonomies (e.g. private/public; biomedical/non-biomedical data, etc.) and the decreased effectiveness of familiar mechanisms such as "consent," "anonymization," and the like. In other words, a new ethical framework has to take an ecosystem perspective and be attentive to the shifts that are currently underway in data-driven research.

References

Almishari, Mishari, Mohamed Ali Kaafar, Gene Tsudik, and Ekin Oguz. 2014. Are 140 characters enough? A large-scale linkability study of tweets. http://arxiv.org/pdf/1406.2746.pdf. Accessed 19 Sept 2015.

Anema, A., S. Kluberg, K. Wilson, et al. 2014. Digital surveillance for enhanced detection and response to outbreaks. *The Lancet Infectious Diseases* 14(11): 1035–1037. doi:10.1016/S1473-3099(14)70953-3.

Angrist, Misha. 2007. *Here is a human being: At the dawn of personal genomics*. New York: Harper.

Auffray, Charles, and Leroy Hood. 2012. Systems biology and personalized medicine–the future is now. *Biotechnology Journal* 7(8): 938–939.

Ayres, J.W., B.M. Althouse, and M. Dredze. 2014. Could behavioral medicine lead the web data revolution? *The Journal of the American Medical Association* 311(14): 1399–1400. doi:10.1001/jama.2014.1505.

Cate, F.H., and V. Mayer-Schönberger. 2013. Notice and consent in a world of Big Data. *International Data Privacy Law* 3(2): 67–73. doi:10.1093/idpl/ipt005.

Christie, G.P., K. Patrick, and D. Schmuland. 2015. Consultation for collective action on personalized health technology: Eliminating ethical, legal, and social barriers for individual and societal benefit. *Journal of Health Communication: International Perspectives* 20(8): 867–868. doi:10.1080/10810730.2015.1063404.

Dawkins, Richard. 1982. *The extended phenotype: The gene as the unit of selection*. Oxford/San Francisco: W.H. Freeman and Company.

Dove, E.S., B.M. Knoppers, and M.H. Zawati. 2014. Towards an ethics safe harbor for global biomedical research. *Journal of Law and the Biosciences* 1(1): 3–51. doi:10.1093/jlb/lst002.

Duhigg, Charles. 2012. How companies know your secrets. *New York Times*, February 16.

Feiler, Bruce. 2014. The United States of metrics. *New York Times*, May 16.
Felch, Jason. 2008. DNA Databases blocked from the public. *Los Angeles Times*, August 29.
Fox, Susanne. 2011. The social life of health information. *Pew Research Center*.http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/. Accessed 19 Sept 2015.
Freifeld, C.C., J.S. Brownstein, C.M. Menone, et al. 2014. Digital drug safety surveillance: Monitoring pharmaceutical products in twitter. *Drug Safety* 37(5): 343–350. doi:10.1007/s40264-014-0155-x.
Gasser, Urs. 2015. Perspectives on the future of digital privacy. *ZSR II* 134: 426–427.
Gasser, Urs, Ryan Budish, and Sarah Myers West. 2015. Multistakeholder as Governance Groups: Observations from case studies. *Berkman Center Research Publication No. 2015–1*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2549270. Accessed 19 Sept 2015.
Gächter, T., and G. Werder. 2015. Gedanken zur allfälligen Verankerung eines «Rechts auf Kopie» in der schweizerischen Bundesverfassung. See also entry in the Swiss Parliament https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20154045.
Ginsburg, G. 2014. Medical genomics: Gather and use genetic data in health care. *Nature* 508: 451–453. doi:10.1038/508451a.
Gleibs, I.H. 2014. Turning virtual public spaces into laboratories: Thoughts on conducting online field studies using social network sites. *Analyses of Social Issues and Public Policy* 14: 352–370. doi:10.1111/asap.12036.
Global Alliance for Genomics and Health. 2015. Privacy and security policy. https://genomicsandhealth.org. Accessed 19 Sept 2015.
Hafen, E., D. Kossmann, and A. Brand. 2014. Health data cooperatives – Citizen empowerment. *Methods of Information in Medicine* 53(2): 82–86. doi:10.3414/ME13-02-0051.
Hayden, E.C. 2015. Genome researchers raise alarm over Big Data. *Nature*. 312–314. doi:10.1038/nature.2015.17912.
Heger, Monica. 2015. Regulators move toward adverse event reporting via mobile apps. *Nat Med* 21: 104. doi:10.1038/nm0215-104.
Homer, Nils, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4(8): e1000167. doi:10.1371/journal.pgen.1000167.
Hood, L., and M. Flores. 2012. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* 29(6): 613–624.
Hood, L., and C. Auffray. 2013. Participatory medicine: a driving force for revolutionizing healthcare. *Genome Med* 5(12): 110.
Jain, S.H., B.W. Powers, J.B. Hawkins, and J.S. Brownstein. 2015. The digital phenotype. *Nature Biotechnology* 33(5): 462–463. doi:10.1038/nbt.3223.
Juengst, E.T., R.A. Settersten Jr., J.R. Fishman, and M.L. McGowan. 2012. After the revolution? Ethical and social challenges in 'personalized genomic medicine'. *Per Med* 9(4): 429–439. doi:10.2217/pme.12.37.
Kahn, J.P., E. Vayena, and A.C. Mastroianni. 2014. Opinion: Learning as we go: Lessons from the publication of Facebook's social-computing research. *Proceedings of the National Academy of Sciences* 111(38): 13677–13679. doi:10.1073/pnas.1416405111.
Koenig, B.A. 2014. Have we asked too much of consent? *Hastings Center* 44(4): 33–34.
Kosinski, M., D. Stillwell, and T. Graepe. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America* 110(15): 5802–5805. doi:10.1073/pnas.1218772110.
Levinson, Daniel R. 2012. Hospital incident reporting systems do not capture most patient harm. Department of health and human services. http://oig.hhs.gov/oei/reports/oei-06-09-00091.pdf. Accessed 17 Oct 2013.
Mandeville, K.L., M. Harris, L.H. Thomas, Y. Chow, and C. Seng. 2014. Using social networking sites for communicable disease control: Innovative contact tracing or breach of confidentiality? *Public Health Ethics.* 7(1): 47–50. doi:10.1093/phe/pht023.

Manson, Neil C., and Onora O'Neill. 2007. *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Markham, Annette, and Elizabeth Buchanan. 2012. Ethical decision-making and Internet research, Recommendations from the AoIR Ethics Working Committee. http://aoir.org/reports/ethics2.pdf. Accessed 19 Sept 2015.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A revolution that will transform how we live, work, and think*. London: John Murray.

Mitchell, C., L.B. Moraia, and J. Kaye. 2014. Health database: Restore public trust in care data project. *Nature* 508: 458. doi:10.1038/508458e.

Mittelstadt, B.D., and L. Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.

de Montjoye, Y.-A., L. Radaelli, V.K. Singh, and A. Pentland. 2015. Unique in the shopping mall: On the reidentifiability of credit card data. *Science* 347(6221): 536–539. doi:10.1126/science.1256297.

Murdoch, T.B., and A.S. Detsky. 2013. The inevitable application of Big Data in health care. *The Journal of the American Medical Association.* 309(13): 1351–1352. doi:10.1001/jama.2013.393.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf. Accessed 19 Sept 2015.

National Institutes of Health. 2014. NIH Genomic Data Sharing Policy. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html. Accessed 19 Sept 2015.

Narayanan, Arvid, and Edward Felten. 2014. No silver bullet: De-identification still doesn't work. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf. Accessed 19 Sept 2015.

Nuffield Council of Bioethics. 2010. *Medical profiling and online medicine: The ethics of 'personalised healthcare' in a consumer age*. London: Nuffield Council on Bioethics.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*, 4–18. London: Nuffield Council on Bioethics.

O'Brien, David, Jonathan Ullman, Micah Altman, Urs Gasser, Michael Bar-Sinai, Kobbi Nissim, Salil Vadhan, Michael John Wojcik, and Alexandra Wood. 2015. Integrating approaches to privacy across the research lifecycle: When is information purely public? *Berkman Center Research Publication No. 2015–7*.http://dx.doi.org/10.2139/ssrn.2586158. Accessed 19 Sept 2015.

Ohm, Paul. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701–1777.

Oliver, J.M., M.J. Slashinski, T. Wang, P.A. Kelly, S.G. Hilsenbeck, and A.L. McGuire. 2012. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics* 15(2): 106–114. doi:10.1159/000334718.

O'Neill, Onora. 2002. *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

O'Neill, Onora. 2013. Can data protection secure personal privacy? In *Genetic privacy: An evaluation of the ethical and legal landscape*, ed. Terry Sheung-Hung Kaan and Calvin Wai-Loon Ho, 25–40. London: Imperial College Press.

Palfrey, John, and Urs Gasser. 2012. *Interop: The promise and perils of highly interconnected systems*. New York: Basic Books.

Paul, Maria. 2015. Your phone knows if you are depressed. *Northwestern News*, July 15.

Pentland, Alex, Todd G. Reid, and Tracy Heibeck. 2013. Big Data and health. Revolutionalizing medicine and public health. *Report of the Big Data and Health Working Group*.http://kit.mit.edu/sites/default/files/documents/WISH_BigData_Report.pdf. Accessed 19 Sept 2015.

Polonetsky, Jules, Omer Tene, and Joseph Jerome. 2014. Benefit-risk analysis for Big Data Projects. Future of Privacy Forum. http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf. Accessed 19 Sept 2015.

Rivers, Caitlin M. and Bryan L. Lewis. 2014. Ethical research standards in a world of Big Data. 3*F1000Research.*

Rudder, Christian. 2014. *Dataclysm: Who we are (when we think no one's looking.)*. UK: Harper Collins. New York.

Saeb, S., M. Zhang, C.J. Karr, et al. 2015. Mobile Phone Sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *J Med Internet* 17(7): e175. doi:10.2196/jmir.4273.

Samaritans. 2014. Samaritans launches Twitter app to help identify vulnerable people. http://www.samaritans.org/. Accessed 19 Sept 2015.

Schneier, Bruce. 2014. *Data and goliath: The hidden battles to collect your data and control your world*. New York: W.W. Norton & Company.

Schwartz, Paul M., and Daniel J. Solove. 2011. The PII problem: Privacy and a New concept of personally identifiable information. *New York University Law Review* 86(2011): 1814.

Secretary's Advisory Committee on Human Research Protections. 2015. Human subjects research implications of "Big Data". http://www.hhs.gov/ohrp/sachrp/commsec/hsrimplicationsofbig_datastudies.html. Accessed 19 Sept 2015.

Sengupta, Somini. 2012. Should personal data be personal? *New York Times*, February 4.

Shaw, Jonathan. 2014. Why "Big Data" is a big deal. *Harvard Magazine*, March-April. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal. Accessed 19 Sept 2015.

Stephens, Zachary D., S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, et al. 2015. Big Data: Astronomical or genomical? *PLoS Biology* 13(8): e1002195. doi:10.1371/journal.pbio.1002195.

Sweeney, Latanya. 2000. Simple demographics often identify people uniquely. http://dataprivacylab.org/projects/identifiability/paper1.pdf. Accessed 19 Sept 2015.

The Economist. 2014. Ebola and Big Data: Waiting on hold. *The Economist*, October 25.

Van Noorden, R. 2014. US agency updates rules on sharing genomic data. *Nature*. doi:10.1038/nature.2014.15800. http://www.nature.com/news/us-agency-updates-rules-on-sharing-genomic-data-1.15800.

Vayena, E., A. Ganguli-Mitra, and N. Biller-Andorno. 2008. Guidelines on biobanks: emerging consensus and unresolved controversies. In *Ethical issues in governing biobanks : global perspectives*, ed. B. Elger, N. Biller-Andorno, A. Mauron, and A.M. Capron, 23–35. Farnham: Ashgate.

Vayena, Effy, and John Tasioulas. 2013. Adapted standards: Ethical oversight of participant–led health research. *PLoS Medicine* 10(3): e1001402. doi:10.1371/journal.pmed.1001402.

Vayena, E., A. Mastroianni, and J. Kahn. 2013. Caught in the web: Informed consent for online health research. *Science Translational Medicine* 5(173): 173fs6. doi:10.1126/scitranslmed.3004798.

Vayena, E., M. Salathé, L.C. Madoff, and J.S. Brownstein. 2015. Ethical challenges of Big Data in public health. *PLoS Computational Biology* 11(2): e1003904. doi:10.1371/journal.pcbi.1003904.

Vayena, Effy, and Urs, Gasser. 2016. Between opness and privacy in genomics. *PLoS Medicine* 13(1): e1001937. doi: 10.1371/journal.pmed.1001937

Watson, Sarah M. 2014. Data is the new "____" on the industrial metaphors of Big Data. http://dismagazine.com/discussion/73298/sara-m-watson-metaphors-of-big-data/. Accessed 19 Sept 2015.

Weber, G.M., K.D. Mandl, and I.S. Kohane. 2014. Finding the missing link for Big biomedical data. *The Journal of the American Medical Association* 311(24): 2479–2480. doi:10.1001/jama.2014.4228.

Wesolowski, Amy, C.O. Buckee, L. Bengtsson, E. Wetter, X. Lu, and A.J. Tatem. 2014. Commentary: containing the ebola outbreak – The potential and challenge of mobile network data. *PLoS Currents Outbreaks*. 2014 Sept 29. doi:10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.

Widdows, Heather. 2013. *The connected self. The ethics and governance of the genetic individual*. Cambridge: Cambridge University Press.

World Health Organization. 2002. Safety of medicines – A guide to detecting and reporting adverse drug reactions – Why health professionals need to take action. http://apps.who.int/medicinedocs/en/d/Jh2992e/12.html. Accessed 19 Sept 2015.

Zimmer, Michael. 2010. "But the data is already public": On the ethics of research in Facebook. *Ethics Information Technology* 12: 313–325. doi:10.1007/s10676-010-9227-5.

Using Transactional Big Data for Epidemiological Surveillance: Google Flu Trends and Ethical Implications of 'Infodemiology'

Annika Richterich

Abstract This chapter provides a critique of methodological developments in influenza surveillance enabled by digital technology. While public health surveillance conventionally relies on data from clinical and virological diagnosis or mortality rate statistics, approaches in 'infodemiology' (Eysenbach, AMIA Ann Symp Proc 244–248, 2006) are based on big data retrieved from Internet sources. Such data indicating the health situation of a population are hence not biomedical data in a traditional sense, since the information may be derived from websites, newswires, or web search logs. After providing an overview of developments in epidemiological surveillance since the 1980s, the chapter discusses Google Flu Trends (GFT) as case study. GFT is an influenza-surveillance application based on web search logs. From November 2008 until August 2015, it was offered by Google Inc. as public 'nowcasting' service with continuous updates. The relevant data are still being collected and provided to selected research institutions, but they are merely presented in retrospect. GFT uses search queries as indicators of influenza-intensities. These queries may be related to a person's medical condition, but they may as well be influenced by external factors such as news coverage. Moreover, the project is based on transactional big data which are exclusively available to Google Inc., and selected academic or governmental institutions. This chapter addresses the implications of such entanglements between public health services, emerging digital technology and corporate objectives. In order to highlight which norms and values are articulated through GFT and to discuss its ethical implications, the chapter employs a pragmatist approach (Keulartz, Sci Technol Hum Val 29(1):3–29, 2004).

A. Richterich (✉)
Faculty of Arts and Social Sciences, Maastricht University, Grote Gracht 90-92, Maastricht, 6211 SZ, The Netherlands
e-mail: a.richterich@maastrichtuniversity.nl

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_3

1 Introduction

The public online service Google Flu Trends (GFT) was developed by Google Inc. researchers in collaboration with the US Centers for Disease Control and Prevention (CDC). From November 2008 until August 2015, it presented influenza-estimations based on users' web search queries and was updated daily (Ginsberg et al. 2008; see Fig. 1). Initially, GFT was developed by relating health data on influenza intensities – publicly provided by the CDC – to selected user search queries. These search queries were linked to topics such as influenza complications or symptoms. Based on correlations between actual influenza intensities and 45 search queries, a web application was created which aimed at estimating "the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day" (Ginsberg et al. 2008: 1012). While GFT has been discontinued as public 'nowcasting'[1] service, the relevant data (web search queries) are still being collected, shared with selected academic and governmental institutions, and published in retrospect.

Since 2013, the service has been repeatedly criticised.[2] During the last years, it has been rightly pointed out that Google Flu Trends illustrates "Traps in Big Data Analysis" and "Big Data Hybris" (Lazer et al. 2014) and that regular miscalculations

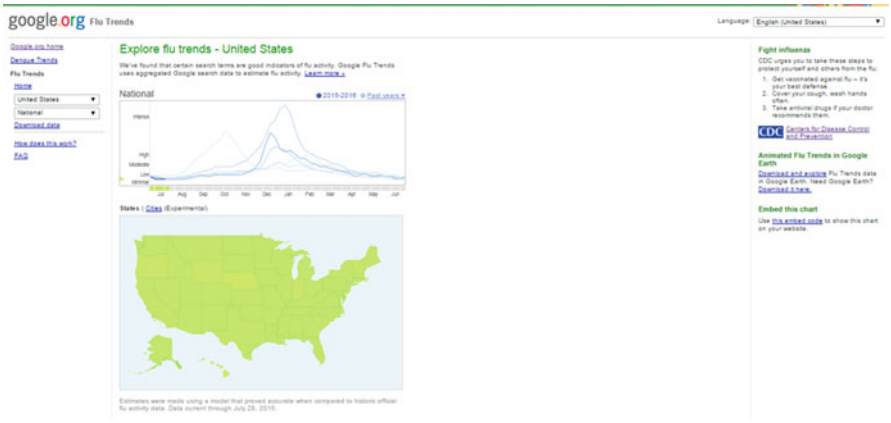

Fig. 1 Screenshot of Google Flu Trends as public 'nowcasting' service for the United States (July 2015). Source: https://www.google.org/flutrends/intl/en_us/us/#US (The service has meanwhile been deactivated.) (© 2015 Google Inc., used with permission. Google and the Google logo are registered trademarks of Google Inc.)

[1]The term 'nowcasting' was used in order to emphasise the minimal delay between data collection, processing and their publication.

[2]See e.g. Butler (2013), Bilton (2013), Lazer et al. (2014), Arthur (2014), Tahir (2014). I will discuss these criticisms in more detail below.

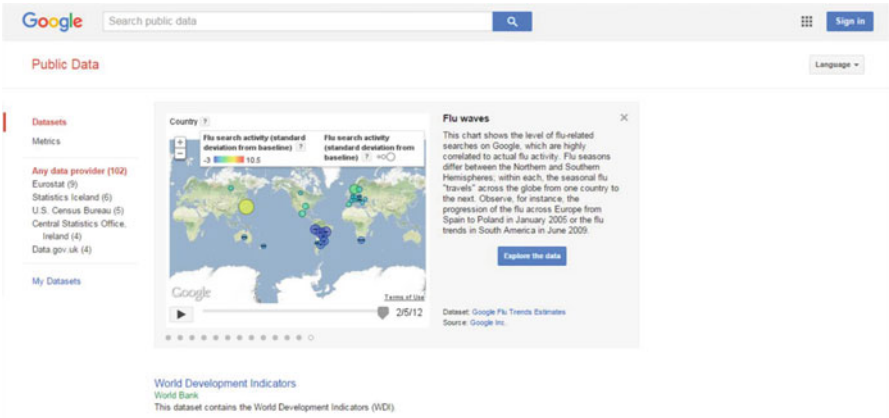

Fig. 2 Screenshot of Google Flu Trends data presented in the *Public Data Explorer* (August 2015). Source: https://www.google.com/publicdata (© 2015 Google Inc., used with permission. Google and the Google logo are registered trademarks of Google Inc.)

should serve as reminder that these services may "complement, but not substitute for, traditional epidemiological surveillance networks" (Butler 2013). However, most critics have focused on the question *if* the service works and how its continuous overestimations may be explained. With their valuable investigations, many of these articles have allowed Google Inc. to refine GFT and to correct flaws. At the same time, these articles and investigations have neglected an important issue: rather than mainly focusing on if the service is functional and to point out reasons for GFT's malfunctioning, it should also be discussed under which conditions it is supposed to work. This perspective is crucial to the following analysis which will highlight ethical implications of emerging influenza-surveillance methods enabled by digital technology and transactional big data.

As mentioned already, GFT has been discontinued as public 'nowcasting' service. Since August 2015, the estimations are not continuously updated, but merely presented as historical overview in Google Inc.'s Public Data Explorer (see Fig. 2).[3] According to the current (August 2015) GFT information website, the real-time data are shared with the Columbia University's Mailman School of Public Health, the Boston Children's Hospital/Harvard, and the CDC Influenza Division. I will continue speaking of GFT in present tense and as ongoing project, since this development shows that the health data collection based on Google web search logs persists, while it is even more difficult now for individuals/institutions which are not authorised by Google Inc. to evaluate these data. It is part of the ethical and

[3] See https://www.google.com/publicdata. The service is available since 2010. It is based on data and forecasts retrieved from (inter)national organisations and institutions which have been subsequently processed and visualised by Google Inc. In the case of GFT, it also involves Google internal data.

methodological dilemma which I will elaborate on in this chapter that we cannot assess in which form and under which conditions these data are shared, processed and used.

GFT is a prominent example of approaches in epidemiological surveillance of influenza, trying to use big data and digital sources in order to predict infectious disease activity in a certain region/population. Already in the late 1990s, independent and corporate researchers have started to explore how internet sources and big data may instruct epidemiological surveillance. As I will show in a historical overview, initial projects mainly employed deliberately published data such as health information provided on websites or via newswires. With the popularisation of search engines, most notably Google, the significance of users' web search queries gained in importance. Such "transactional data (the traces of people online behavior)" (Manovich 2011, p.12) appeared to be insightful in predicting infectious disease intensities. Already in 2006, Eysenbach illustrated in an independent study that Google search queries could be used in order to estimate influenza-intensities in Canada (see Eysenbach 2006, p.244ff.). Two years later, Google Inc. launched the public service Google Flu Trends and Ginsberg et al. (2008) published the first article on its creation.[4]

This chapter will analyse these developments and particularly the case study GFT by drawing on a pragmatist approach to ethics as suggested by Keulartz et al. (2004). Moreover, I will emphasise the interplay between promises and risks (e.g. unreliability and privacy concerns) as highlighted by Rip (2013). While I will explain this approach in more detail in the following section, these are the main questions derived from my methodological framework:

- 'Traditional ethics': What kind of normative assumptions, which arguments, justifications and values are articulated and neglected in the (corporate) presentation, scientific and public debates of Google Flu Trends?
- 'Discourse ethics': What kind of institutional interdependencies emerge through services such as Google Flu Trends? How do such relations define scientific and public debates and facilitate certain constellations of knowledge production and knowledge control (i.e. power)? On the one hand, corporations now dominate access to certain health-relevant data (based on web search logs) and they likewise control what these data may be used for. On the other hand, we should not only look at Google web search log data, but also at the publicly available health data which were used to substantiate the service: who should be allowed to use these governmental data, for what purposes?

The chapter is divided into three main sections: an elaboration on the pragmatist approach to ethics used in this chapter, a brief historical overview of influenza-

[4]From May 2011 until August 2015, the company also offered *Google Dengue Trends*. The service documented the likeliness of dengue epidemics for countries such as Bolivia, India, Indonesia or Singapore (http://www.google.org/denguetrends). The development process has been described in a paper by Chan et al. (2011); see also Gluskin et al. (2014).

surveillance, and a case study of Google Flu Trends (structured according to the abovementioned questions). The historical overview of influenza-surveillance starts in the 1980s, but is mainly focused on developments since the 2000s. I will discuss differences between traditional and recent surveillance approaches, particularly with regards to the types of (big) data retrieved and the institutions involved. I will elaborate on recent studies and approaches with regards to two main developments: these services are characterised by fundamental changes in the data collection; consequently, these methodological transitions come along with a different nature of the biomedical (big) data itself.

Subsequently, the chapter will discuss Google Flu Trends as case study. The influenza-monitoring application will be explained in more detail. I will show which values, arguments and justifications were decisive for the public presentation of the service and how it was taken up in scientific and public discussions. I will question what kind of understanding of online privacy is promoted through GFT. Moreover, I will critically assess under which conditions it produces and presented its calculations. In the analysis of the institutional context ('discourse ethics'), I will emphasise the users' position and the ethical implications inscribed in such (potentially) health-relevant data which are exclusively available to respective media companies, their advertising customers and selected scientists/institutions.

In conclusion, I will argue that Google Flu Trends was supposed to be staged as philanthropic investment and corporate 'data philanthropy' – however, the service unintentionally also indicated that it was only one out of many big data mining results based on the fact that users automatically pay their search engine queries with the data they leave behind. It moreover drew attention to the methodological challenges and uncertainties attached to the use of emerging digital technologies and big data for epidemiological surveillance and health research more generally. I will point out methodological and ethical implications of entanglements between emerging technologies, the corporations controlling these technologies as well as the data they produce, and public health institutions.

2 A Pragmatist Approach to Ethics

In order to highlight the ethical implications of epidemiological surveillance more generally and GFT in particular, I will draw on Keulartz et al.'s suggestions for a "pragmatist approach to ethics in a technological culture" (2004, p.14). The authors develop this methodological framework not as "a complete alternative for other forms of applied ethics but rather a complement", aimed at a "new perspective on the moral and social problems and conflicts that are typical for a technological culture" (Keulartz et al. 2004, p.5).[5] Their approach combines the

[5]For a broader contextualisation of pragmatist ethics see Keulartz et al. (2002), LaFollette (2000), Joas (1993).

methodological, empirical strengths of science and technology studies (STS) with the normative lens of applied ethics. By combining these research fields, the authors intend to counterweight respective weaknesses. According to Keulartz et al., applied ethics used to have "insufficient insight into the moral significance of technological artifacts and systems" and "it therefore cannot cope adequately with the dynamic character of our technological culture" (2004, p.25). Their pragmatist approach hence draws on insights from STS, particularly with regards to an acknowledgement of technology's inherent agency, its influence on humans and their environments. At the same time, it employs aspects of applied ethics in order to overcome a (back then) blind spot in STS: its reluctance to address and make normative statements.

This chapter will mainly focus on the *context of justification* of GFT and epidemiological surveillance based on transactional big data. Drawing on Keulartz et al.'s framework, I will first provide a product-focused analysis of arguments, justifications, values and norms which are articulated in GFT, i.e. how it was presented by Google Inc. and in public debates (scientific journals and their impact on news media). Secondly, Keulartz et al. propose the use of *discourse ethics* (Apel 1988; Habermas 1990, 1994).[6] This method will instruct the second part of my case study. It aims at aiding fair processes of public debate, deliberation and decision making: "to develop procedures and institutions that guarantee equal access to public deliberation and fair representation of all relevant arguments to ensure that moral decisions are based on the 'force of the better argument' rather than on the force of power, money, and the like" (Keulartz et al. 2004, p.19). With regards to discourse ethics (in order to show the relations between affected actors and institutions), I will address the following questions:

- What kind of power dynamics are in play and shape the possibilities for debate and technology development in a particular way ('institutional context')?
- Which actors are currently involved in the debate, which actors are neglected, but should be involved ('stakeholder analysis')?

In addition, my approach was inspired by Rip's concept of "pervasive normativity".[7] In his investigation of emerging technologies and constructive technology assessment, the author stresses the relevance of "strategy articulation of actors involved in emerging technologies, and to recognize that there are normativities

[6]In addition to these methods which are aimed at the *context of justification*, the authors suggest addressing the *context of discovery* with the product-focused approach of "dramatic rehearsal" and the process-focused "conflict management" (Keulartz et al. 2004, p.19). These options should be seen as part of a 'toolbox' however, and are not conditional elements to be covered in such a pragmatist approach to ethics. Instead, the authors "propose that depending on the moral problem at hand, pragmatists will switch between these different tasks and their corresponding methods or tools" (ibid: 18).

[7]Rip's conceptualises "pervasive normativity" as approach "in the spirit of pragmatist ethics, where normative positions co-evolve" (2013, p.205). He claims however that his concept addresses a potential shortcoming: "Pragmatist ethics may be able to capture the normativities involved, but has to overcome its micro-level focus" (Rip 2013, p.205).

involved that may actually contribute to ethics (as pragmatist ethics would emphasise) (Rip 2013, p.196). Rip treats normativity as anthropological category: as an inherent feature of social life and practices, rather than as a qualifier which would allow us to differentiate between 'normative' and 'non-normative' (ibid, p.192). He pays particular attention to the promises and fears articulated in the field of emerging technologies. His perspective has been particularly insightful with regards to the promises related to GFT.

3 Historical Overview

Epidemiology, the science of patterns and factors related to public health and disease conditions, has undergone significant changes since the 1980s. Most recently, these are related to technological developments such as the popularisation of digital media and emerging possibilities to access and analyse vast amounts of global online user data. Epidemiological surveillance is a sub-discipline of epidemiology. It involves systematic, continuous data collection, documentation and analysis of information which reflects the current health status of a population.[8] It aims at providing a reliable information basis for governments, public health institutions and professionals to react adequately and timely to potential health threats. Ideally, epidemiological surveillance enables the establishment of early warning systems for epidemic outbreaks in a geographic region or even pandemics (multinational/global).

Main sources relevant to traditional epidemiological (public health) surveillance are mortality data, morbidity data (case reporting), epidemic reporting, laboratory reporting, individual case reports and epidemic field investigation (see Declich and Carter 1994). The data sources may vary however depending on the development and standards of a country's public health services and medical facilities. Since the 1980s at the latest, computer technology and digital networks have become increasingly influential factors – not merely with regards to archiving and data analysis, but in terms of communication and exchange between relevant actors and institutions. As Declich and Carter pointed out in a section of their paper called "Ways to improve the system" [of public health surveillance]:

> The introduction of computer networks is opening a completely new way of performing traditional surveillance activities. The main advantage of networking is improved data timeliness that allows better monitoring of diseases and rapid identification of epidemics and changing epidemiological patterns. The quick return of information to the data collectors, together with access to on-line information, can stimulate participation. There are two well-described experiences with computer networks: one in the USA, the Epidemiologic Surveillance Project which links weekly reporting of notifiable infectious diseases from

[8] Until the mid twentieth century, the term 'surveillance' was used in medical contexts in order to refer to the health status of an individual person (e.g. with regards to necessary quarantine). It was only after the 1950s that the term was used for referring to the spreading of a particular disease within a certain population (Reintjes and Krämer 2003, p.57).

> State Health Departments to the CDC via computer; and one in France, the French Communicable Disease Network, initiated in November 1984, which includes the National Department of Health and local health offices with part of the transmission from the local to the national level occurring through the network (Dechlich and Carter 1994, p.299).

Dean et al. (1994) depicted a similar vision of new approaches in epidemiological surveillance in a chapter subsection "Overview of a surveillance system in the future". To them, particularly the improved connection and communication between medical experts and affected individuals seemed crucial: "Ideally the epidemiologist of the future will have a computer and communications system capable of providing management information on all these phases and also capable of being connected to individual households and medical facilities to obtain additional information" (Dean et al. 1994, p.246). In this sense, advances in the field of digital information and communications technology have been commonly seen as chances for improvements in epidemiological surveillance.

The aforementioned *French Communicable Disease Network*, with its *Réseau Sentinelles*, was a decisive pioneer in computer-aided approaches. It was one of the first, systematic attempts to build a system for public health/epidemiological surveillance based on computer networks. Meanwhile, it may seem almost self-evident that the collected data are available online. Weekly/annual reports present intensities (ranging from "minimal – very high activity") for 14 diseases, including 11 infectious diseases such as influenza.[9] Similar (public) services are provided by the *World Health Organisation*'s (*WHO*) "Disease Outbreak News",[10] the "Epidemiological Updates"[11] of the *European Centre for Disease Prevention and Control* (*ECDC*) or (merely for influenza cases in Germany and during the winter season) by the *Robert Koch Institute's* "Consortium Influenza". With its *Project Global Alert and Response* (GAR), the *WHO* additionally establishes a transnational surveillance und early-warning system. It aims at creating an "integrated global alert and response system for epidemics and other public health emergencies based on strong national public health systems and capacity and an effective international system for coordinated response".[12]

The use of digital technology and digital data processing has hence significantly affected approaches in epidemiological surveillance since the 1980s. However, one aspect remained unchanged: these approaches still focused on detecting actual cases of illness and diseases. The respective technology is mainly used in order to improve the communication and processing of biomedical data. These data are based on quantifications of actual disease diagnoses. Such 'traditional' systems of epidemiological surveillance are based on biomedical data which are digitally

[9] See https://websenti.u707.jussieu.fr/sentiweb/?site=fr

[10] See http://www.who.int/csr/don/en/index.html

[11] See http://ecdc.europa.eu/en/press/epidemiological_updates/Pages/epidemiological_updates.aspx

[12] See http://www.who.int/csr/en/

documented, analysed and presented. Therefore, these systems rely on computer-aided approaches for processing biomedical data, but the health data are not 'digitally native' themselves.

3.1 Infodemiology: Covering 'Supply' and 'Demand'

While these 'traditional' approaches in epidemiological surveillance rely on data from clinical and virological diagnosis or mortality rate statistics, more recent studies aim at analysing big data retrieved from Internet sources. During the early 1990s, such methods were mainly targeted at publicly available information. Eysenbach (2002, 2006, 2009) described this intersection between epidemiology and digital information and communications technology as *infodemiology* or *infoveillance*.[13] Since then, the understanding of the term has changed corresponding to digital technology developments.

Early publications in this field (until the mid-1990s) discussed the availability, distribution and quality of health information provided and potentially accessed by affected individuals online. In 2002, Eysenbach still defined *infodemiology* as "the study of the determinants and distribution of health information and misinformation" (p.763). In this context, he related the term to studies which highlighted the quality of medical information on topics such as diabetes (Davison 1996) or fever (Impicciatore et al. 1997). Four years later, he broadened his definition – also in the light of his own approach, the "Google ad sentinel method":

> [T]he development of 'infodemiology' metrics based on automated tracking and analysis of the distribution and determinants of health information (both supply and need) in a population and/or information space is possible and can provide important clues and evidence for public health policy and practice. In a broader sense, an 'infodemiology' science is needed to develop a methodology and real-time measures (**indices**) **to understand patterns and trends for general health information**, [. . .] **and to understand the predictive value of what people are looking for** (**demand**) for syndromic surveillance and early detection of emerging diseases (Eysenbach 2006, p.247) [emphasis added].

This differentiation and expansion of the field is mainly induced by certain technological developments and possibilities. While most scholars focused on the side of 'information supply/indices', the popularisation of search engines as well as platforms which allowed for a direct articulation of users have led to an increasing significance of assessing users' demand for health information. One can hence determine a methodological development which evolves from an analytic focus on 'supply' data to 'demand' data regarding *re-/enquired* health information.

[13] Eysenbach was not the only one to suggest terms describing the emerging field. For example, Breton et al. (2012) coined the term *epimining* in order to refer to quantitative analyses of certain terms used in online sources/media which, according to the authors, "allows the extraction of information from web news (based on pattern research) and a fine classification of these news into various classes" (p.1)

Looking at the field of *infodemiology* today, one can differentiate between the following approaches; epidemiological surveillance based on:

- 'professional', public information online (e.g. *Health Map*, *Global Public Health Intelligence Network*)
- explicit, conscious information provided by users and affected individuals (e.g. *Flu Near You*, *Grippeweb*)
- implicit information provided (mainly) unconsciously by users (e.g. Google Flu Trends; *Yahoo* research/Polgreen et al. 2008)

Already in 1997, the *WHO* and the *Health Canada's Centre for Emergency Preparedness and Response* (*CEPR*) were working on a prototype for the (subscription-based) *Global Public Health Intelligence Network* (*GPHIN*; Health Canada 2003). In 2005, Mawudeku and Blench described GPHIN as "unique multilingual system" which "gathers and disseminates relevant information on disease outbreaks and other public health events by monitoring global media sources such as news wires and web sites" (p.9).

Such 'supply'-oriented approaches can still be found in more recent studies. For example, Breton et al. (2012) showed that semantic analyses of sources such as the *Agence France-Presse* (*AFP*) may be used for predicting epidemic intensities. Likewise, not only professional media communication may be useful, but also the analysis of social media communication shows potential. Chunara, Andrews and Brownstein (2012) employed data from the microblogging-platform *Twitter* in order to detect the outbreak of cholera in Haiti and to monitor the intensity of the epidemic. With regards to this data source, it remains unclear to what extent one is dealing with information which has been provided 'consciously' by users, in the sense that they were aware of a further use of their tweets. In addition, this approach made use of data derived from the *HealthMap* project[14] which is based on an automatic analysis of semantic content from blogs, news-websites, RSS feeds as well as official surveillance data.

Services such as *Flu Near You*[15] or the German *Grippeweb*[16] (transl. 'Flu Web'; *Robert Koch Institute*) pursue crowdsourcing strategies. They rely on the conscious participation of volunteers providing information on their own health status or

[14] See http://healthmap.org/en. The service as been developed by researchers affiliated with the *Boston Children's Hospital* and the *Harvard Medical School* (Brownstein et al. 2008). It combines different data sources, such as *Twitter* feeds and platforms such as Google news, with official public health reports. While it has been launched in 2006 already, it has received most media attention since the Ebola outbreak in 2014. On March 14, the site first picked up on news reports about a hemorrhagic fever, while the *WHO* only officially reported on the Ebola outbreak more than a week later (March 23, 2014). In this instance, it was of course not classified as Ebola yet, but the information could have acted as early indicator.

[15] See https://flunearyou.org/. The service is closely related to *HealthMap*, and has been developed by epidemiologists from *Harvard University* and the *Boston Children's Hospital* as well as the *The Skoll Global Threats Fund*.

[16] See https://grippeweb.rki.de/

(potential) influenza symptoms. These approaches depend on a deliberate effort of volunteers and their faithful, correct information. Due to the limited scope of this chapter, I will not be able to discuss the aforementioned approaches in more detail. Particularly the involvement of volunteers appears to be an interesting development which emphasises users' deliberate, conscious involvement rather than their automatic 'mining' for health relevant information. However, these approaches raise issues regarding users' capability for self-diagnosis and sincerity in participation.

3.2 Analysing Health Information Demand

This chapter focuses on approaches in epidemiological surveillance drawing on transactional big data: the documentation and analysis of user behaviour (i.e. their search queries) with regards to health relevant information. Big data, produced by the search terms entered through vast amounts of users worldwide, form the basis for these attempts. In particular, studies by Eysenbach (2006), Polgreen et al. (2008) and Ginsberg et al. (2008) have explored this aspect of *infodemiology*. They all start from the assumption that certain search queries may be motivated by influenza or influenza-like-illness (ILI), either experienced by the individual her-/himself or in her/his social environment. Assuming that a certain search query correlates (steadily) with actual influenza intensities, it may be used as indicator of disease dynamics.

Initially, Eysenbach explored this research field with his *Google ad sentinel method*. He was able to demonstrate "an excellent correlation between the number of clicks on a keyword-triggered link in Google with epidemiological data from the flu season 2004/2005 in Canada" (Eysenbach 2006, p.244). Eysenbach described his approach as a "trick" (ibid, p.245), since the actual Google search queries were not available to him. Hence, he had to create a *Google Adsense* commercial campaign in order to obtain the necessary data. His method was not able to obtain actual search query quantifications, but only allowed him to factor in those users who subsequently clicked on a presented link. When (Canadian) Google users entered "flu" or "flu symptoms", they were presented with an ad *Do you have the flu*? created by Eysenbach. The link led to a health information website regarding influenza. As an (alleged) advertising customer, Google Inc. provided him with quantitative information and geographic data. When relating these to data from the governmental *FluWatch Reports* (*Public Health Agency Canada*), he detected a positive correlation between the increase of certain search queries and influenza activities.

In 2008, Polgreen et al. presented a similar study design. The researchers, one of them from *Yahoo Research*, were provided with data from *Yahoo Inc*. web search logs. Based on queries related to influenza (March 2004-May 2008) and internet protocol addresses which allowed for their geographic localisation, the researchers created a database which was then related to the data of traditional surveillance

systems.[17] Just like Eysenbach, they asserted a correlation between certain search terms and actual influenza-intensities. Apart from emphasising the cost-efficient advantages of their approach, the researchers highlighted that predictions could be calculated in a very timely manner: "With use of the frequency of searches, our models predicted an increase in cultures positive for influenza 1–3 weeks in advance when they occurred" (Polgreen et al. 2008, p.1443).

When Ginsberg et al. published their results in November 2008,[18] they hence did not present a completely new approach. However, their publication was accompanied by the launch of a public Google Inc. service[19] in 2008. Former studies had merely emphasised the methodological potential of web search queries. The authors summarise their investigation: "Because the relative frequency of certain queries is highly correlated with the percentage of physicians visits in which a patient presents influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day" (Ginsberg et al. 2008, p.1012).

4 Case Study: Google Flu Trends

The aforementioned studies are all enabled by the fact that digital user activities such as the use of search engines are not ephemeral, but are turned into transactional big data. Ginsberg et al. used data provided by Google Inc.'*s* market leading search engine: they were hence derived from databases documenting users' search queries. As a result, the GFT interface (July 2015) illustrated historic and estimated influenza intensities in geographic maps as well as line graphs (see Fig. 1). The service is not merely a result of analysing web search queries: during its development phase (and for its adjustment), the researchers had to draw on biomedical data provided by traditional epidemiological surveillance networks. In addition to the search queries data, they used two main data sources. They employed data publicly provided by the *CDC* for nine U.S. surveillance regions as well as state-reported ILI percentages for Utah. The *CDC* publish information regarding the amount of patients which

[17] Mainly two data sources were relevant for this project: "Each week during the influenza season, clinical laboratories throughout the United States that are members of the World Health Organization Collaborating Laboratories or the National Respiratory and Enteric Virus Surveillance System report the total number of respiratory specimens tested and the number that were positive for influenza. The second type of data summarize weekly mortality attributable to pneumonia and influenza. These data are collected from the 122 Cities Mortality Reporting System" (Polgreen et al. 2008, p.1444)

[18] The paper was originally published online on November 19, 2008, but was corrected on February 19, 2009 (see http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html#cor1).

[19] Strictly speaking, GFT is part of *Google.org*, a Google Inc. initiative (see also Strom and Helft 2011).

have been diagnosed with influenza or 'influenza-like-illness' (ILI) online (see http://www.cdc.gov/flu/weekly). During flu/influenza season, these data are updated weekly.

Ginsberg et al. retrieved these data and tested them for correlations with selected search queries. They developed a database of potentially relevant queries which were subsequently related to the data provided by the CDC: "For the purpose of our database, a search query is a complete, exact sequence of terms issued by a Google search user [. . .]. Our database of queries contains 50 million of the most common search queries [. . .]" (Ginsberg et al. 2008, p.1014). Originally, this database consisted of "hundreds of billions of individual searches from 5 years [2003–2008] of Google web search logs" (ibid, p.1012). The top 45 queries showing a correlation with increasing influenza/ILI intensities were then chosen as initial basis for the construction of GFT. These search queries were allegedly related to topics such as influenza complications and symptoms or certain antibiotic medication. However, as Lazer et al. pointed out, the exact terms have never been disclosed and moreover "the examples that have been released appear misleading" (2014, p.1204). This is already indicative for GFT's tendency to 'black-box' certain information which is on the one hand crucial in order to understand its functioning, but may on the other hand facilitate miscalculations. In the following two sub-sections, I will analyse GFT with regards to such issues by drawing on the pragmatist perspective which I outlined before.

4.1 Normative Assumptions, Justifications and Values

First, I will highlight the normative assumptions – the arguments, justifications and values –which are articulated and neglected in the (corporate) presentation, scientific and public debates of Google Flu Trends. When GFT was initially presented in a paper titled "Detecting influenza epidemics using search engine query data" (Ginsberg et al. 2008), the authors emphasise advantages and promises, but likewise pointed to conditions and risks of GFT. The paper starts with an affirmation of obvious threats posed by influenza epidemics: the illnesses and deaths causes by seasonal influenza epidemics as well as the incalculable health threat of new strains of influenza virus. With regards to these risks, the authors claim to have developed a model implemented in the service GFT which estimates influenza activity with a reporting lag of one day (and is hence considerably quicker than traditional influenza surveillance networks which provide data with a reporting lag of 1–2 weeks). The authors summarise this in a later section of the paper:

> Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today. Whereas traditional systems require 1-2 weeks to gather and process surveillance data, our estimates are current each day (Ginsberg et al. 2008, p.1014).

As I explained before, this model relies on monitoring the health-seeking behaviour of users represented by their queries in the online search engine Google. One main condition for its functionality is hence a sufficiently large population of search engine users. Likewise, it is based on cooperation with the *CDC*, using the influenza data which are publicly accessible online. One needs to keep in mind that these data are merely provided for the influenza seasons. Therefore, the model used for GFT could only involve data describing ILI activities for these time frames. As I will explain below, it has been pointed out that this approach was most likely also responsible for early miscalculations in GFT. The cooperation during the development is described as continuous process of 'sharing' with the *Epidemiology and Prevention Branch of the Influenza Division* at the *CDC* to assess its timeliness and accuracy (see ibid, p.1013). Hence the *CDC* served as source of validation in order to ensure the accuracy of the data. Despite the abovementioned claims regarding improved efficiency and timeliness, the authors describe GFT not as solitary service, but as initial indication for further responses to potential epidemics. The system is not suggested as "replacement for traditional surveillance" (ibid.); instead, these influenza estimations are meant to "enable public health officials and health professionals to respond better to seasonal epidemics" (ibid, p.1013). GFT is hence not supposed to estimate and predict influenza in an isolated way: it is offered as knowledge source and early warning system to be used by health professionals. On www.cdc.gov/flu/weekly, GFT is mentioned (above the *WHO* and *Public Health Canada/England*), however it remains unclear to what extent and how these data were/are in fact used by the *CDC* (or other health professionals). While the authors describe GFT mainly as information tool instructing the decision making and responses of health professionals and institutions, the public version of the service seems to neglect this aspect: it suggests itself as public information source for ILI intensities.

Despite presenting GFT as tool instructing further strategies and investigations, the authors also anticipated a main source of miscalculations: users' search engine queries may not only be triggered by individual health conditions, but may also be influenced by e.g. news about geographically distant influenza outbreaks. Hence, the dynamics of users' search engine behaviour act as potential confounders of data used to instruct GFT. This connection highlights two issues: the service is susceptible to "Epidemics of Fear" (Eysenbach 2006, p.244); moreover, it relies on users which are ideally not influenced by any other knowledge despite their own health condition or experiences in their immediate, social environment.

4.1.1 Epidemics of Fear

Already in 2006, Eysenbach advised caution with regards to the significance of web search queries, since they may "be confounded by 'Epidemics of Fear'" (Eysenbach 2006, p.244). Also the developers of GFT pointed out this possibility:

> In the event that a pandemic-causing strain of influenza emerges accurate and early detection of ILI percentages may enable public health officials to mount a more effective early response. Although we cannot be certain how search engine users will behave in such a scenario, affected individuals may submit the same ILI-related search queries used in our model. Alternatively, panic and concern among healthy individuals may cause a surge in the ILI-related query fraction and exaggerated estimates of the ongoing ILI percentage (Ginsberg et al. 2008, p.1014).

These concerns draw attention to the fact that the motivations for users to enter certain search queries may vary over time. While a certain query may initially indicate a person's individual illness, it may later be influenced by influenza activities in different areas or even countries. Since these differently motivated search queries would be automatically fed into the GFT model, this would lead to miscalculations. It is hence vulnerable to deviations in users' behaviour. The search queries which were assessed as significant and originally showed a positive correlation might suddenly *mean* something different. The algorithm which is ultimately crucial to the calculation of influenza estimates is hence prone to miscalculations caused by changes in users' motivations to enter certain search queries. This methodological uncertainty relates to the conditions of big data retrieval: while the data are continuously produced, their usage context and conditions are highly dynamic. Certain queries which are identified and used as 'health data' are in fact only temporary indicators which may turn into influenza interest or health concern data (without actually signifying a person's health condition). One should not assume that certain search queries may function as consistent, indexical sign.

In fact, this concern was confirmed several times: for example, in 2009, accompanying the H1N1 virus, as well as in the beginning of 2013, GFT calculations by far overestimated actual influenza intensities as indicated by the *CDC*. As Butler pointed out in his article "When Google got flu wrong" (2013) the algorithms defining GFT results need to be continuously adjusted to the dynamic user behaviour in order to avoid miscalculations:

> [T]he latest US flu season seems to have confounded its algorithms. Its estimate for the Christmas national peak of flu is almost double the CDC's (see 'Fever peaks'), and some of its state data show even larger discrepancies. It is not the first time that a flu season has tripped Google up. In 2009, Flu Trends had to tweak its algorithms after its models badly underestimated ILI in the United States at the start of the H1N1 (swine flu) pandemic—a glitch attributed to changes in people's search behaviour as a result of the exceptional nature of the pandemic (Butler 2013).[20]

Hence, relying on GFT data is particularly risky in cases which could cause deviations in users' search behaviour. It requires ongoing adjustment and evaluation, since any deviation from historically assessed search patterns may act

[20]Moreover, a study funded by *Google.org*, in cooperation with the *CDC*, referred to influenza intensities which were not predicted as part of the model which mainly relied on seasonal influenza patterns: "The 2009 influenza virus A (H1N1) pandemic [pH1N1] provided the first opportunity to evaluate GFT during a non-seasonal influenza outbreak. In September 2009, an updated United States GFT model was developed using data from the beginning of H1N1" (Cook et al. 2011).

as confounder. This is especially crucial, since such deviations are difficult to predict.[21] The service had been adjusted subsequently to the raised issue. However, the implications and risks for health professionals and institutions considering the service as source of health information remained. Miscalculations are particularly likely in cases in which new, external factors influence the motives which are crucial to search queries considered to be 'influenza-/ILI-relevant'. While the algorithms 'assume' certain motivations, these may have changed. In order to correct these misinterpretations, the GFT data again need to be related to data provided by traditional surveillance networks. This also means that an assessment of eventual errors is at best as fast as those systems. Brownstein, who is also involved in the aforementioned collaborative influenza project *Flu Near You* commented on this condition for GFT in Butlers article: "You need to be constantly adapting these models, they don't work in a vacuum [...] You need to recalibrate them every year" (Brownstein quoted in Butler 2013).

Consequently, GFT can be described as application which needs to be constantly 'work-in-progress'. The data need to be reassessed with the use of traditional influenza health data in order to adjust the algorithms. The service depends on a continuous data evaluation which reassesses the relation between relevant search queries and assumed (health/influenza) motivations for entering those. The frequent overestimations of GFT are therefore caused by the fact that search queries are a big data source which constantly and unpredictably changes its meaning. From an ethical perspective, these insights are relevant in several ways: first of all, they imply certain assumptions about the ideal user providing data for GFT; secondly, they question to what extent health professionals and institutions may rely on such data; lastly, one also needs to consider that the interplay described before does not only allow *Google.org* to improve GFT, but also to understand users' search behaviour in relation to current developments more generally. Moreover, for Google Inc. it became quickly clear that GFT could not be maintained without continuous investments and required a certain expertise. Seeing the severe criticisms which were raised, these investments certainly did not pay off in terms of positive publicity. Hence, the recent deactivation as nowcasting service and the delegation of the data assessment to public health professionals and (academic) institutions also shows that such projects can be unsustainable and volatile due to the corporate interests involved. These aspects will be addressed in the following sections.

[21] In the abovementioned case, Butler stated that "[s]everal researchers suggest that the problems may be due to widespread media coverage of this year's severe US flu season, including the declaration of a public health emergency by New York State last month. The press reports may have triggered many flu-related searches by people who were not ill" (2013, p.156).

4.1.2 The 'Innocent User' as Ideal Data Source

As I have illustrated above, the data retrieved for Google Flu Trends are meant to indicate the health situation of a population. Instead of being derived from virological or clinical diagnosis however, they are based on search queries. This also means that they may be related to a person's medical condition, but they may as well be triggered by external factors such as media coverage of influenza. The ideal source for GFT is hence an 'innocent user' who is largely uninfluenced by external factors and whose knowledge does not disrupt the algorithms crucial to determining influenza-intensities. Of course, this assumption is only applicable to the extent that GFT is unable to pick up on such deviant user behaviour – which can be seen as its ultimate challenge.

As indicated above, GFT is particularly prone to miscalculations caused by (unconscious or deliberate) deviations in user motivations for selected search queries. The service is hence currently based on the fact that users enter search queries, before actually knowing anything more specific about the wider societal circumstances of their health condition. This fact also facilitates an interest in a certain non-transparency of the service and a 'black-boxing' of its functional conditions. As Lazer et al. criticised, the Google/*CDC* researchers have been quite unspecific with regards to the search terms selected for GFT. In fact, they even seemed to be misleading (Lazer et al. 2014, p.1204). One reason for this lack of disclosure may also be that a publication of exact search queries could in turn influence the frequency of these terms. As Arthur (2014) described, by drawing on Lazer et al.'s paper, already functions such as *Google Autocomplete* may have unintentionally encouraged and increased certain search queries which were significant to GFT (Fig. 3).

4.1.3 Privacy

While the exact search queries have never been disclosed, Ginsberg et al. also assured in their paper that:

> [N]one of the queries in the Google database for this project can be associated with a particular individual. The database retains no information about the identity, internet protocol (IP) address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being made anonymous in accordance with Google's privacy policy (http://www.google.com/privacypolicy.html) (Ginsberg et al. 2008, p.1014).

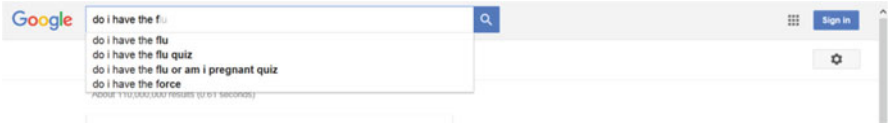

Fig. 3 Screenshot of Google's *Autocomplete* function. Source: www.google.com (© 2015 Google Inc., used with permission. Google and the Google logo are registered trademarks of Google Inc.)

In this sense, privacy is allegedly ensured by prohibiting the possibility to obtain information which is directly related to a user's identity/name (after 9 months). This was also implied on the former service website itself, where users were informed:

> Your personal search data remains safe and private. Our graphs are based on aggregated data from millions of Google searches over time. Moreover, the results Google Flu Trends displays are produced by an automated system (https://www.google.org/flutrends/intl/en_gb/about/how.html; source is no longer accessible).

The amount of user queries and their aggregation are used as argument for ensuring privacy. In addition, the reference to "an automated system" suggests that the information is not actually read and exploited. This may of course be true in the sense that it may not be read and analysed explicitly by humans; however, the automated process nevertheless extracts information and derives conclusions about users behaviour (see Schermer 2011). What remains neglected here is the fact that such data collection approaches enable the construction of user profiles which can be employed to address users' with 'relevant', i.e. potentially profitable information such as advertisement. In this sense, services such as GFT promote a concept of privacy which is in fact not adequate for the era of big data (see e.g. Hildebrandt and Koops 2010; Leese 2014). This seems particularly relevant since the philanthropic purpose of GFT itself provides a strong normative justification and serves as legitimisation of such an understanding of privacy. While the strictly commercial products of Google Inc. are more vulnerable to privacy claims, GFT sets user privacy off against the greater good represented by epidemiological surveillance.

It seems symptomatic in this context that early privacy concerns raised by non-profit organisations such as the *Electronic Privacy Information Center* and *Patient Privacy Rights* were dismissed as "misplaced nagging" (Madrigal 2014), since they seemed to misjudge the handling of users' data. Instead – and I will get back to this point in the stakeholder analysis – GFT representatives as well as public health organisations endorsing such services need to face the question whether the service calls attention to issues regarding user privacy which are not covered with our pre-established conceptions of privacy. This is an issue which is also closely related to user-consent and an active clarification of GFT's conditions for functioning. While users are informed about basic functionalities of the service when visiting the GFT website, most Google search engine users are oblivious to the various uses of their transactional data for this particular service or Google's advertisement programmes. The relation between GFT and Google Inc.'s commercial programmes will be discussed in more detail in the following section "Corporate Entanglements".

4.2 Discourse Ethics

In this section on discourse ethics, I will discuss the institutional conditions which define the emergence, maintenance and debate of GFT. This will comprise an analysis of the institutional context preceding and enabling the service as well as a stakeholder analysis.

4.2.1 Institutional Context

When considering ethical implications of GFT, one needs to acknowledge the institutional power dynamics which shape how the technology has been developed, debated and eventually modified (see Friedman and Nissenbaum 1996). In this context, particularly questions of data access and information disclosure are relevant. I already indicated in the initial explanation of *discourse ethics* that such a perspective is concerned with societal dynamics which may facilitate or inhibit "equal access to public deliberation and fair representation of all relevant arguments" (Keulartz et al. 2004, p.19). Hence, one needs to address how the corporate embedding of GFT influences the possibilities for its public assessment. It is crucial to understand discourse ethics in the context of Habermas' "theory of communicative action" (see Mittelstadt et al. 2015, p.11). According to Habermas, we have to assume that any human communication – such as the initial presentation of GFT as well as its subsequent negotiation – poses certain *validity claims* regarding *truth*, (normative) *rightness* and *authenticity/sincerity*.[22] As Mittelstadt et al. explain, "[c]ommunicative action requires the speaker to engage in a discourse whenever any of these validity claims are queried. This implies a willingness to engage with the interlocutor, to take her seriously and to be willing to change one's position in the line of that argument (...)" (2015, p.11). In case of GFT, already the fundamental possibility to assess these validity claims seems considerably constrained: I have already shown that critics of GFT have e.g. questioned if the correct information has been provided by its developers, if the service is based on correct assumptions about users and its own technological conditions, what it means for users' privacy and what kind of corporate interests may be implied in GFT. Hence, validity claims concerning the service, stated by Google Inc. or rather its representatives, have been challenged. One of the main problems seems to be however that the lack of methodological transparency prohibits a factual assessment of crucial validity claims. Their public contestation is restricted to the level of speculations. Due to their enormous scope and the variety of projects in which Google Inc. is involved, a reaction as described with the concept of communicative action is extremely difficult to achieve. While the company initially reacted with adjustments of the service, a critical engagement with the public opinion was largely missing (e.g. limited to research blog posts). Ultimately, the decision to discontinue GFT as public nowcasting service indicates that the public contestation of the service has led to a communicative strategy on the side of the company which allows for even less insights into the use of search query data for purposes such as influenza surveillance. This also shows that (unintentionally) neglecting certain actors and stakeholders has created discourse conditions limited by asymmetrical power relations, in this cases

[22] Due to the limited scope of this chapter, Habermas' theory will not be explained in detail, but only partially with respect to those aspects relevant to the following argumentation. For more extensive accounts on the relevance of discourse ethics for emerging technologies see Mittelstadt (2013) and Mingers/Walsham (2010).

defined by restricted information and data access. Therefore, the following sections depict the relations between Google Inc./*Google.org* and researchers involved in GFT, external/independent researchers evaluating the service and the users who contribute to its functioning with their transactional data.

Data Hierarchies and Monopoly

The paper published by Ginsberg et al. (2008) only discloses certain information about the development of the service. The search queries, but also exact quantities and the relevant algorithms have never been revealed. Hence, it leaves external/independent scientists guessing about certain functionalities, implications and conditions of the development process. At the same time, the service relies on the publicly available data provided by health institutions such as the *CDC*. Transactional big data such as Google search queries are only accessible to Google Inc./*Google.org* and selected researchers. While they enable services such as GFT, they are also crucial for the company's business model. Disclosing certain data would be on the one hand problematic in terms of users' privacy, on the other hand it would render the data useless (or rather: costless) for advertising purposes. In academic contexts, the exclusiveness of these data has facilitated approaches such as a reverse engineering of the initial algorithms, suggested by Lazer et al. (2014), as well as the initially mentioned 'trick' employed by Eysenbach. As Manovich criticises, it is characteristic for transactional big data that they are only accessible to corporations and selected (industry) partners: "Only social media companies have access to really large social data – especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not" (2011, p.5). This constellation implies asymmetrical power relations between Google Inc./*Google.org* and external researchers as well as institutions, due to Google's data monopoly.

The actual data are exclusively available to Google Inc. and selected partners or customers. The public and external researchers are meanwhile dealing with strategically limited indicators. These are presented in a way that they convey certain information, without actually disclosing the underlying data. While certain companies enable scientists and the public to download their big data via open application programming interfaces (see Manovich 2011, p.5ff.), Google Inc. produces indicators which allow for an assessment of relations and intensities, but does neither offer any numerical nor semantic accuracy. Needless to say, the same goes for the algorithms.[23] GFT is largely a 'black box' and the level of actual big data remains opaque for the public, including external scientists.

[23] See also Ippolita (2013, p.75ff).

As a result, researchers have only limited possibilities to assess GFT. This seems especially problematic in the light of a main challenge discussed with regards to big data. As boyd points out, data retrieval may be easier than ever for certain actors. However, the data analysis becomes likewise more and more problematic:

> Social scientists have long complained about the challenges of getting access to data. Historically speaking, collecting data has been hard, time consuming, and resource intensive. Much of the enthusiasm surrounding Big Data stems from the opportunity of having easy access to massive amounts of data with the click of a finger. Or, in Vint Cerf's [since 2005 vice president and Google Inc.'*s* 'chief internet evangelist', A.R.] words, 'We never, ever in the history of mankind have had access to so much information so quickly and so easily.' Unfortunately, what gets lost in this excitement is a critical analysis of what this data is and what it means (boyd 2010).

The exclusive access to data, controlled by *Google.org* and hence ultimately Google Inc., also inhibits the academic assessment of GFT and the validity claims posed by the service. While it has immensely profited from comments by external academics (whose criticisms were used in order to revise the algorithms and the calculation model), these external assessments remain to some extent speculative. The conditions for such valuable external evaluations are hence far from ideal.

Lazer et al. argue that the overestimations (e.g. during the 2011–2012 flu season) are caused by so-called "big data hybrids" (2014, p.1203) and are errors which could have been avoided or at least corrected. According to the authors, particularly the combination of big data and 'small data' was problematic; the researchers were trying "to find the best matches among 50 million search terms to fit 1152 data points" (ibid.). Hence, there was high chance "that the big data were overfitting the small number of cases – a standard concern in data analysis" (ibid.). The lack of methodological disclosure however inhibits such investigations (and possibly also an improvement of the service). Moreover, the necessary recalibration of algorithms had only been implemented very few times for GFT as public 'nowcasting' service (after the H1N1 pandemic in 2009, in October 2013 and in October 2014, see Stefansen 2014) – which raises doubts about the sustainability of the approach.

At the same time, one should not only look at Google web search log data, but also at the publicly available health data which are used to substantiate the service: considering such new utilisations, who should be allowed to use these public health data, for what purposes? In order to develop and refine the service, GFT depends on data regarding actual influenza intensities provided by institutions such as the *CDC* or the *ECDC*. Hence it benefits from the public availability of these data. As a service, GFT emphasises how influenza-intensities may be derived from web search logs and hence what can be learned from them. It de-emphasises however what the corporation itself may learn about its own data by accessing publicly available health data. For the company and eventually its advertising customers, GFT also sheds light on users' search logics and motivations. Therefore, it may serve as valuable lesson in understanding potential customers.

Corporate Entanglements

Google Flu Trends is not an isolated service. It should be seen as part of Google Inc.'s big data portfolio. Similarly, *Google Trends* allows users to explore general search query intensities and *Google Correlate* indicates relations between certain queries.[24] Those search queries which are identified as influenza/ILI relevant may likewise be of interest for commercial customers from e.g. the pharmaceutical industry. Even more importantly, GFT is based on search queries entered in Google. Therefore, economically oriented changes to the search engine – such as the aforementioned introduction of *Google Autocomplete* – may also influence the GFT results. Corporate interests potentially interfere with the service. It is hence not only the user behaviour which is dynamic and may act as confounder of GFT estimations. Moreover, Google itself might encourage different behaviour and search results. Lazer et al. (2014) rightly refer to the possibility that changes in search behaviour may be due to "'blue team' dynamics – where the algorithm producing the data (and thus user utilization) has been modified by the service provider in accordance with their business model" (p.1204). Such issues show that the embedding of a declared non-profit health service into a corporate, digital platform leads to entanglements with corporate interests which can confound its functionality. Apart from the corporate entanglements, one is moreover left speculating how Google ensures that health relevant and hence highly sensitive data are in the long term protected from e.g. insurance companies' or governmental access.

4.2.2 Stakeholder Analysis

This section looks at some of the actors and institutions who are currently involved in the debate as well as those groups who are neglected, but should be involved.

Google Inc. and Google.org

GFT is presented as part of Google Inc.'*s* non-profit branch *Google.org*. The service is enabled however by data which are collected as part of corporate services. The company has hence access to certain health-relevant data (in this case based on web search logs) and likewise controls what these data may be used for. The company selects the scientists involved in research concerning these data, and it defines to what extent governmental institutions such as the *CDC* may have access. At the same time, it draws on publicly available health data in order to develop and maintain GFT. As outlined above, the emerging data monopolies and inhibited possibilities for an external academic assessment of the service are hardly reflected upon. While GFT is suggested as tool to instruct the work of public health professionals, these do in fact have very little possibilities to evaluate this suggested basis for their work and potentially far reaching decisions. The slow evaluation and

[24] See Mohebbi et al. (2011).

adjustment processes raise further doubts concerning the reliability of the service. It is unclear to what extent *Google.org* sees GFT as sustainable, long-term project serving the health sector and acknowledges certain responsibilities, or if it was rather an attempt to explore the possibilities of big data. This questions seems especially warranted since the most recent developments suggest that Google Inc. has transferred the responsibility for analysing the GFT data to (very few) selected academic and governmental institutions.

Moreover, with regards to the users whose data are used for GFT, privacy issues and user consent are largely neglected. The service itself used to comment on privacy issues, but the search engine Google does not address the broader question for what kind of purposes users' data may be used. The use of the search engine is treated as automatic consent to documenting and analysing the emerging transactional big data – e.g. as it is now continued with the transmission of search query data to certain institutions.

External/Independent Researchers

While there have been various (aforementioned) attempts by 'Google-external' researchers to shed light on GFT functions, these were largely based on deductions derived from public Google services and were hence partly speculative. Independent researchers have insufficient access in order to assess the data. Data access becomes therefore a privilege which is attached to either the employment by Google Inc./*.org* or the selective contracting of certain researchers. The described data hierarchies and disparity result in a situation in which the development of the GFT model remains partly a black box. Just like the data access, also the adjustments of GFT are controlled by *Google.org*. It is unclear which criteria are crucial for an adjustment of the algorithm, i.e. under which conditions does *Google.org* deem it vital to adjust GFT. We do not know which search terms were decisive for GFT over time, or how their anonymisation was ensured. Such conditions also prevent researchers from assessing validity claims regarding truth as well as (in consequence) rightness. While the lack of transparency may be explained by referring to user privacy (this of course also has a strategic aspect), other reasons are GFT entanglements with corporate interests as well as its dependence on the 'innocent user'. Overall, external researchers have limited to no access to the GFT data and the relevant algorithms. The conditions for an eventual access remain unclear. However, the notification on GFT's discontinuation as public service now refers to a form for a "Research Interest Request".[25]

[25]The form states: "Use this form to request access to Google Flu Trends and Google Dengue Trends signals for research and nowcasting purposes. Please note that access will be granted only to selected research partners" (https://docs.google.com/forms/d/1_I0bALRi3kWRcWppj-OtrojZGb9Wbwpz40Q669oSbS8/viewform?rd=1).

It seems questionable on which level external researchers should analyse GFT. Strategies of reverse engineering are surely insightful (but also even more difficult due to the recent changes in the service). However, in addition to addressing the (mal-) functioning of GFT, actors such as researchers and public health professionals need to reflect on the data hierarchies and the bias produced through such services. While many authors have in fact pointed out valuable criticism, they have done so under rather difficult conditions, caused by asymmetrical power/knowledge relations.

Governmental/Public Health Institutions

The role of public health institutions is at least twofold: on the one hand, they are encouraged to use GFT as public health indicator, instructing further measures. Moreover, data provided by the *CDC* and the *ECDC* served as basis for the development of GFT in various countries. Therefore, one needs to raise the questions: Who defines how they support the development of such services? How should public health professionals and institutions evaluate, react to and use data provided by a service like GFT?

The promises of such services seem somewhat overemphasised, leading to exaggerated hopes in the new data sources for epidemiological surveillance. Search engine queries are presented as rather 'natural, uninfluenced' data which are e.g. unbiased by users' shame to ask a physician or pharmacist. In similar contexts, this kind of allegedly unbiased data has led to a certain excitement also among researchers. Already with regards to privacy issues concerning the automated analysis of electronic data exchange, the authors of a *nature* editorial commented: "For a certain sort of social scientist, the traffic patterns of millions of e-mails look like manna from heaven" (nature editorial board 2007, p.637). This reaction seems now reproduced in the context of epidemiological surveillance. In 2006, Larry Brilliant, former director of Google.org said in an interview with wired magazine: "I envision a kid (in Africa) getting online and finding that there is an outbreak of cholera down the street. I envision someone in Cambodia finding out that there is leprosy across the street" (Zetter 2006).

While GFT has been explicitly presented as less ambitious in the paper, the service itself did not make the limited possibilities explicit. The hopes towards such services were also reflected in statements of public health professionals: "'Social media is here to stay and we have to take advantage of it,' says Taha Kass-Hout, Deputy Director for Information Science at the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia" (Rowland 2012). This quote implies that drawing on big data provided by digital media is per-se advantageous and moreover presents a stable data source. Both assumptions are problematic. First of all, it is uncertain how respective services providing the data will change and/or how the users will change the behaviour leading to such data. On the one hand, certain information may confound the search patterns relevant to GFT; on the

other hand, broader concerns such as state or corporate surveillance may lead to changes in user behaviour. Such data sources are hence by no means stable – even if they continuously produce data, these may be motivated differently and need to be constantly assessed. More generally, Google Inc.'s most recent decision to discontinue GFT as public nowcasting service also shows that the service as such will be subject to changes. Therefore, public health professionals need to remain critical towards approaches such as GFT, since neither their accuracy nor their sustainability is warranted.

Secondly, this is not only a methodological problem, but results in ethical concerns regarding the utilisation of such big data: What kind of problems are we facing, when the products of corporations overlap and merge with programmes relevant to public health communication and disease prevention? boyd summarised this issue: "Just because it is accessible doesn't mean using it is ethical" (2010; see also boyd and Crawford 2012). Public health institutions and professionals also have a responsibility to reflect on these problems: they need to consider the implications of cooperation with companies such as Google Inc. and also the use of their own data for certain purposes. (As I have pointed out before, one should not be misled by the fact that GFT is a *Google.org* product, since the data used in this context are part of Google Inc.'*s* big data portfolio.) Public health institutions need to discuss the fact that health-relevant data are now owned and controlled by an international corporation rather than governmental institutions. In addition, the publicly provided data by these institutions support the development of services such as GFT and hence an understanding of search engine user behaviour in a larger context. As a 'side-effect', the use of publicly available health data for the development of GFT may also facilitate an understanding of user behaviour: how users react to certain events, how search queries may change over time, and how search engine functions may encourage respective terms.

Search Engine Users and User Representatives

Shortly after the release of GFT, Rotenberg (*Electronic Privacy Information Center*) and Peel (*Patient Privacy Rights*) sent an open letter to Eric Schmidt, former Google Inc. CEO, expressing concerns regarding potential privacy threats. More specifically, the authors requested:

> Would you agree to publish the technique that Google has adopted to protect the privacy of search queries for Google Flu Trends? As you know, there is considerable debate as to what constitutes 'anonymized' data [. . .] If Google has found a way to ensure that aggregate data cannot be re-identified, it should publish its results (Rotenberg and Peel 2008).[26]

[26]The scepticism implicit in this last sentence may also be explained in the context of early attempts to release allegedly anonymised web search logs. For example, the *AOL* publication of users' search queries has shown how difficult their actual anonymisation may be (see Arrington 2006).

It has already been described with regards to external researchers that GFT is largely a 'black box', but of course, this also accounts for the users enabling the service. Concerns as stated by Rotenberg and Peel have remained largely uncommented, except for the standard reference to Google's privacy policy. The launch of GFT also came along with unreflective polemics (by external commentators) regarding privacy concerns, such as: "Yet keeping search logs for 9 months may be useful for dealing with advertising-related questions and for optimizing a search engine's responsiveness. If users don't like that, nobody's forcing them to use Google" (McCullagh 2008).[27] Later on, Rotenberg and Peel's privacy concerns were called "misplaced nagging" (Madrigal 2014).

Instead of using Google's market dominance as authoritative argument for users' acceptance of any use of their data, debates (under conditions allowing for a factual assessment of validity claims and involving communicative action) as well as legal regulations are needed which address possibilities for informing users and facilitating control over their data. Currently, users have no possibility to opt out of the utilisation of their transactional data. It is however questionable who should own those data and who should be able to access them and control their use. Of course, technically this access is defined and limited by Google Inc., but the limited options for users to control and to understand the use of their data is highly problematic. Especially, the lack of legal frameworks for user protection does require reflection on ethical concerns which may substantiate future measures. These debates should be influenced by users, researchers, public health professionals and institutions, and require an increased transparency of Google Inc. regarding the functionalities of GFT and its entanglements with profitable, corporate elements. Moreover, GFT specifically is an initiative which justifies the use of transactional big data with a philanthropic purpose and hence creates a biased impression of the company's use of web search logs. A normative assumption running through all the outlined debates is hence the implicit argument that concerns regarding e.g. privacy or methodological transparency should be neglected for the greater good of epidemiological surveillance, allegedly represented by GFT. In this sense, especially in its early stage, the service implicitly also aimed at promoting the social acceptance of transactional big data uses.

5 Conclusion

Seeing the arguments and normative implications discussed so far, the following main issues are crucial for an ethical evaluation of GFT from a pragmatist perspective. First of all, with regards to the institutional context, the service creates asymmetrical power/knowledge relations due to data hierarchies and a Google

[27] The article ends with the remark: "Disclosure: The author is married to a Google employee" (McCullagh 2008).

monopoly of data access. This inhibits an independent assessment of validity claims regarding truth, rightness and authenticity/sincerity. The debate surrounding GFT is on the one hand influenced by a lack of methodological disclosure, including e.g. the search queries, data quantities and algorithms. This forces scientists to speculate about possible reasons for GFT's frequent overestimations. This initial tendency seems even more distinct since GFT has been discontinued as public nowcasting service in August 2015: Individuals who are not authorised by Google Inc. have even less possibilities to gain insights into the estimations derived from public search queries. These are merely presented in hindsight, as historical overview, in Google Inc.'s so-called *Public Data Viewer* (Fig. 2). Due to the lack of methodological insights, also ethical concerns – e.g. with regards to an anonymisation of web search data – cannot be addressed with reasonable certainty.

While the service's deactivation highlights issues of data accessibility and data collection transparency, it also hints at the strategic interests behind its initial presentation as public service. Briefly after its launch, criticisms regarding GFT's utilisation of transactional big data were more generally compromised by depicting them as disproportionate in the light of the promises of such new approaches in epidemiological surveillance. However, such promises are frequently overstated; a tendency which has been described by Rip:

> [. . .] promises about an emerging technology are often inflated to get a hearing. Such exaggerated promises are like confidence tricks and can be condemned on bordering at the fraudulent. But then there is the argument that because of how science and innovation are organised in our societies, scientists are almost forced to exaggerate the promise of their envisaged work in order to compete for funding and other resources (2013, p.192/193).

In fact, GFT was repeatedly criticised for its overestimations and instability caused by changing user behaviour and Google's internal adjustments. In this context, it is crucial to keep in mind however that the promises of GFT were not only aimed at ensuring further funding, but they also served as justification for the use and exploration of transactional big data. The philanthropic purpose and hopes were utilised as normative argument to justify that corporations and selected scientists may draw on big data and may employ methods which are largely opaque. In this regard, GFT served as promotional flagship of transactional big data and corporate data philanthropy. The data which are framed as 'health data' are however highly context-dependent. While they may be used as data source regarding health-relevant information, they could be likewise used for an analysis of various other purposes, e.g. users' potential demand for medication or other products. The deactivation as public 'nowcasting' service and its reduction to a retrospective data service can also be seen as an attempt of Google Inc. to regain control about the conclusions which observers may draw from their evaluation. The possibility to access the data in the *Public Data Explorer* suggests that the data are being shared, while users in fact do not see the actual data, but the results from black-boxed processes and calculations which have been applied to the relevant search queries. As public nowcasting service, GFT seems to have lost its value for Google Inc., since it is not able to maintain its initial image as philanthropic project tapping into big data, but instead

meanwhile stands for the severe methodological challenges coming along with big data research. While GFT has always been a service which only disclosed a marginal part of the information and data which have been decisive for its development and maintenance, this condition is now all the more applicable.

Particularly the methodological problems have been pointed out quite rightly in various assessments by Google-external researchers who identified recurring reasons for the service's overestimations and unreliability. While these evaluations were conducted under restraints caused by GFT's lack of methodological disclosure, they were able to ascertain that malfunctions were often related to corporate entanglements of the service with changes in the Google search engine (which is responsible for the data collection). The deactivation of GFT as public 'nowcasting' service makes such efforts even more difficult – or almost impossible. Hence, the service's assessment is now increasingly exclusive and restricted by Google Inc.'s authorisation of selected researchers.

As critics have shown, the GFT data collection is prone to confounders caused by internal changes of the Google search engine. At the same time, the dynamic user behaviour is a potential source of miscalculations. Changes in search queries may be influenced by events and influenza information provided by news media, but it is also not excluded that individuals deliberately adjust their behaviour. It may occur that "research subjects (in this case Web searchers) attempt to manipulate the data-generating process to meet their own goals, such as economic or political gain. [. . .] Ironically, the more successful we become at monitoring the behaviour of people using these open sources of information, the more tempting it will be to manipulate those signals" (Lazer et al. 2014, p.1204). Apart from e.g. a political motivation to manipulate GFT, it is also possible that the usage of their transactional data may inhibit users' willingness to ask for health data. As pointed out in the open letter by *EPIC* and *PPR*, this "could also have a chilling effect on Internet users who may be reluctant to seek out important medical information online if they are concerned that their search histories will be revealed to other" (Rotenberg and Peel 2008).

Secondly, the stakeholder analysis highlighted that there is a negligence of relevant actors and (in turn) a negligence of certain responsibilities by involved stakeholders. The users whose Google search queries create the database for GFT can obtain general information on "Privacy" and "Terms". However, if they do not take the initiate, users remain oblivious of the collection and uses of their data, one of the being GFT. Since the service is prone to miscalculations caused by changes in users' motivations for certain search queries, it relies on subjects who are ideally as little as possible influenced by factors which are not related to their own influenza/ILI condition or illness in their direct social environment. The ideal data source for GFT is hence an 'innocent user' who is largely unaffected by wider information concerning a potential influenza epidemic or any other information which might act as confounder with regards to service relevant key words. Such a user image seems rather problematic.

Nevertheless, this may also be considered as one reason for the lack of transparency and methodological reticence. One needs to keep in mind though that these

data are also part of Google Inc.'*s* advertisement programmes and that a disclosure might not only jeopardize the functioning of GFT, but would also render these data costless and hence unprofitable. Besides, it is also possible – despite all claims for a full anonymisation – that the data publication would infringe users' privacy.

Apart from implications for the users, Google'*s* data monopoly inhibits an assessment through external researchers. As described above, the service remains largely a 'black box' and academics interested in an assessment are forced into strategies of reverse engineering and speculation. While Google Inc. itself is highly reluctant when it comes to sharing data, GFT was enabled by the fact that public health institutions such as the U.S. *CDC* and the *ECDC* publish their health data regarding influenza intensities online. What is hence missing are not only ethical and legal guidelines for the (fair) use of transactional big data which need to be taken into account by companies, but also public health institutions need to reflect on how their data are being used. The disparity in public data disclosure in the case of GFT, also raises the question under which conditions such public data should be used and what kind of support should be granted for projects such as GFT. When looking at developments such as Google Inc.'s *Public Data Explorer* – which draws on data retrieved from e.g. the *U.S. Census Bureau*, *Eurostat* and the *Organisation for Economic Co-operation and Development* (OECD) – it becomes obvious that public data are processed by corporations in ways which may not have been foreseen when they were originally published as datasets. This is not merely an issue which applies to health data, but it is particularly problematic due to the sensitivity of such information.

With the emergence of new, technologically facilitated possibilities for epidemiological surveillance of diseases such as influenza, public health professionals are required to carefully assess such new options with regards to ethical affordances. In this context, one needs to keep in mind that exaggerated promises and hopes do not merely function as facilitators for one, philanthropic type of data utilisation. Instead, the use of transactional big data for a philanthropic, non-profit cause such as influenza surveillance also promotes the value of big data approaches more generally.[28] Search engines such as Google opt users into their conditions for use: when wanting to use the search engine, they do not have any other option than paying their search engine queries with the data they leave behind. These may be used for projects such as GFT, but such big data are likewise the core of Google Inc.'s business model. GFT offers merely a glimpse into corporate data mining strategies, company's access and possible utilisations of health-indicative big data. Currently, we are facing new dimensions of data mining, in scope and

[28]Epidemiological surveillance, as an approach which seems obviously dedicated to the greater good, appears to be a popular field for corporate engagement. More recently, *Microsoft Research* presented their "Project Premonition". The initiative aims at employing drones and robotic mosquito traps in order to detect pathogens in mosquitos, i.e. ideally prior to any human infection (see http://research.microsoft.com/en-us/um/redmond/projects/projectpremonition).

with regards to digital methods. At the same time, the access to and control of health-relevant big data shifts from public institutions to corporations – fostering asymmetric power/knowledge relations which inhibit an independent assessment of validity claims.

References

Apel, Karl-Otto. 1988. *Diskurs und Verantwortung: Das Problem des Übergangs zur postkonventionellen Moral*. Berlin: Suhrkamp.

Arrington, Michael. 2006. *AOL proudly releases massive amounts of private data*. TechCrunch. http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data. Accessed 9 Aug 2015.

Arthur, Charles. 2014. Google Flu Trends is no longer good at predicting flu, scientists find. *The Guardian*. https://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu. Accessed 9 Aug 2015.

Bilton, Nick. 2013. Disruptions: Data without context tells a misleading story. *The New York Times*, February 24. http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context. Accessed 9 Aug 2015.

boyd, danah. 2010. *Privacy and publicity in the context of big data*. keynote www 2010, Raleigh, North Carolina. http://www.danah.org/papers/talks/2010/WWW2010.html. Accessed 9 Aug 2015.

boyd, danah., and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society* 15(5): 62–679.

Breton, Didier, et al. 2012. Mining web data for epidemiological surveillance. In *Proceedings of the 2012 Pacific-Asia conference on emerging trends in knowledge discovery and data mining*, ed. Takashi Washio and Jun Luo, 11–21. Berlin/Heidelberg: Springer.

Brownstein, John S, et al. 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med* 5(7). http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050151. Accessed 9 Aug 2015.

Butler, Declan. 2013. When Google got flu wrong. US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature International Weekly Journal of Science* 494: 155–156.

Chan, Emily H, et al. 2011. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *Plos One*. http://www.plosntds.org/article/info%3Adoi%2F10.1371%2Fjournal.pntd.0001206. Accessed 9 Aug 2015.

Charles, Arthur. 2014. Google Flu Trends is no longer good at predicting flu, scientists find. *The Guardian*. http://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu. Accessed 9 Aug 2015.

Chunara, Rumi, et al. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene* 86(1): 39–45.

Cook, Samantha, et al. 2011, Assessing Google Flu Trends performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *Plos One*. www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0023610. Accessed 9 Aug 2015.

Davison, Karen. 1996. The quality of dietary information on the World Wide Web. *Journal of Dietetic Practice and Research* 57: 137–141.

Dean, Andrew G., et al. 1994. Computerizing public health surveillance systems. In *Principles and practice of public health surveillance*, ed. Steven Teutsch and Elliott Churchill, 245–270. New York: Oxford University Press.

Dechlich, S., and Anne O. Carter. 1994. Public health surveillance: Historical origins, methods and evaluation. *Bulletin of the World Health Organization* 72: 285–304.

Eysenbach, Gunther. 2002. Infodemiology: The epidemiology of (Mis)information. *American Journal of Medicine* 113: 763–765.

Eysenbach, Gunther. 2006. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. *AMIA Annual Symposium, Proceedings*: 244–248. http://www.ncbi.nlm.nih.gov/pubmed/17238340.

Eysenbach, Gunther. 2009. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research* 11(1). http://www.jmir.org/2009/1/e11. Accessed 9 Aug 2015.

Friedman, Batya, and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14(3): 330–347.

Ginsberg, Jeremy, et al. 2008. Detecting influenza epidemics using search engine query data. *Nature: International Weekly Journal of Science* 457(7232): 1012–1014.

Gluskin, Rebecca T, et al. 2014. Evaluation of internet-based dengue query data: Google dengue trends. *PLOS Neglected Tropical Diseases* 8(2). http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0002713. Accessed 9 Aug 2015.

Habermas, Jürgen. 1990. Discourse ethics: Notes on a program of philosophical justification. In *Moral consciousness and communicative action*, 43–115, Cambridge, MA: MIT Press.

Habermas, Jürgen. 1994. *Justification and application: Remarks on discourse ethics*. Cambridge: MIT Press.

Health Canada. 2003. *The global public health intelligence network*. http://www.hc-sc.gc.ca/ahc-asc/pubs/_intactiv/gphin-rmisp/index-eng.php. Accessed 9 Aug 2015.

Hildebrandt, Mireille, and Bert-Jaap Koops. 2010. The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review* 73(3): 428–460.

Impicciatore, Piero, et al. 1997. Reliability of health information for the public on the world wide web: Systematic survey of advice on managing fever in children at home. *British Medical Journal* 314: 1875.

Ippolita. 2013. *The dark side of Google (theory on demand 13)*. Amsterdam: Institute of Network Culture. http://issuu.com/instituteofnetworkcultures/docs/tod_13_ippolita_binnenwerk-def-sp?e=3130431/5095074. Accessed 9 Aug 2015.

Joas, Hans. 1993. *Pragmatism and social theory*. Chicago: University of Chicago Press.

Keulartz, Jozef, et al. 2002. *Pragmatist ethics for a technological culture*. Dordrecht: Kluwer.

Keulartz, Jozef, et al. 2004. Ethics in technological culture: a programmatic proposal for a pragmatist approach. *Science, Technology & Human Values* 29(1): 3–29.

LaFollette, Hugh. 2000. Pragmatic ethics. In *The Blackwell guide to ethical theory*, ed. Hugh LaFollette, 400–419. Oxford: Blackwell.

Lazer, David, et al. 2014. The parable of Google Flu: Traps in big data analysis. *Scienc*e 343: 1203–1205. http://j.mp/1ii4ETo. Accessed 9 Aug 2015.

Leese, Matthias. 2014. The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European union. *Security Dialogue* 45(5): 494–511.

Madrigal, Alexis. 2014. In defense of Google Flu Trends. *The Atlantic*, March 27. http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688. Accessed 9 Aug 2015.

Manovich, Lev. 2011. Trending: The promises and the challenges of big social data. www.manovich.net/DOCS/Manovich_trending_paper.pdf. Accessed 9 Aug 2015.

Mawudeku, Abla, and Michael Blench. 2005. Global Public Health Intelligence Network. http://www.mt-archive.info/MTS-2005-Mawudeku.pdf. Accessed 9 Aug 2015.

McCullagh, Declan. 2008. Privacy groups target Google Flu Trends. Cnet. http://www.cnet.com/news/privacy-groups-target-google-flu-trends. Accessed 9 Aug 2015.

Mingers, John, and Geoff Walsham. 2010. Toward ethical information systems: the contribution of discourse ethics. *MIS Quarterly* 34(4): 833–854.

Mittelstadt, Brent D. 2013. *On the ethical implications of personal health monitoring: A conceptual framework for emerging discourses*. Leicester: De Montfort University.

Mittelstadt, Brent D., Bernd C. Stahl, and N. Ben Fairweather. 2015. *How to shape a better future? Epistemic difficulties for ethical assessment and anticipatory governance of emerging technologies. Ethical theory and moral practice ethical theory and moral practice*, 1–21. Dordrecht: Springer.

Mohebbi, Matt, et al. 2011. Google correlate whitepaper. www.google.com/trends/correlate/whitepaper.pdf. Accessed 9 Aug 2015.

Nature editorial board. 2007. A matter of trust: Social scientists studying electronic interactions must. *Nature International Weekly Journal of Science* 449(7163): 637–638. www.nature.com/nature/journal/v449/n7163/pdf/449637b.pdf. Accessed 9 Aug 2015.

Polgreen, Philip, et al. 2008. Using internet searches for influenza surveillance. *Clin Infect Dis* 47(11): 1443–1448.

Reintjes, Ralf, and Alexander Krämer. 2003. *Infektionsepidemiologie*. Berlin/Heidelberg: Springer.

Rip, Arie. 2013. Pervasive normativity and emerging technologies. In *Ethics on the laboratory floor*, ed. Simone van der Burg and Swierstra Tsjalling, 191–212. Basingstoke: Palgrave Macmillan.

Rotenberg, Marc, and Deborah Peel. 2008. Open letter to Google Inc./Eric Schmidt. https://epic.org/privacy/flutrends/EPIC_ltr_FluTrends_11-08.pdf. Accessed 9 Aug 2015.

Rowland, Katherine. 2012. Epidemiologists put social media in the spotlight. *Nature International Weekly Journal of Science*. http://www.nature.com/news/epidemiologists-put-social-media-in-the-spotlight-1.10012. Accessed 9 Aug 2015.

Schermer, Bart W. 2011. The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

Stefansen, Christian. 2014. Google Flu Trends gets a brand new engine. October 31. http://googleresearch.blogspot.nl/2014/10/google-flu-trends-gets-brand-new-engine.html. Accessed 9 Aug 2015.

Strom, Stefanie, and Miguel Helft. 2011. Google finds it hard to reinvent philanthropy. *The New York Times*, January 29. http://www.nytimes.com/2011/01/30/business/30charity.html?_r=0. Accessed 9 Aug 2015.

Tahir, Darius. 2014. Google Flu Trends update shows underlying issues of big data. *Modern Healthcare*, November 3. http://www.modernhealthcare.com/article/20141103/BLOG/311039995. Accessed 9 Aug 2015.

Zetter, Kim. 2006. *Brilliant's wish: Disease alerts*. http://archive.wired.com/science/discoveries/news/2006/02/70280?currentPage=all. Accessed 9 Aug 2015.

Denmark at a Crossroad? Intensified Data Sourcing in a Research Radical Country

Klaus Hoeyer

Abstract Denmark is regularly portrayed in international science journals as 'the epidemiologist's dream': a country where health data on all citizens can be combined with e.g. information about social or financial position, kinship ties, school performance data as well as tissue samples. Moreover, it can all be done without the informed consent of the individual. This chapter describes the practices in Denmark involved in what I call 'intensified data sourcing'. I define intensified data sourcing as attempts at getting more data, of better quality, on more people – and I point out how intensified data sourcing has emerged as a new way of running the health services. My key point with this chapter is that though research uses of health data receive the most attention, research is not necessarily the main purpose with intensified data sourcing. Nevertheless, ethical debates tend to focus on research and thereby neglect an adequate understanding of the everyday practices of data sourcing and the many competing purposes it serves. Furthermore, I point out how ethical debates often focus on the rights of the individual, though data sourcing operates at the level of the population, and when attending to individual rights there is an unfortunate tendency to conjure concerns about privacy with rights of autonomy. We need new modes of ethical reasoning that take point of departure in an understanding of actual data practices. Since Denmark is in many ways at the forefront of intensified data sourcing, it is a good place from which to begin rethinking the policy challenges associated with intensified data sourcing at both national and European levels.

1 Introduction

Routine healthcare activities today facilitate population-based data sourcing on a scale unanticipated just 10 years ago. Thanks to new information and communication technology, patients produce healthcare data of great administrative value

K. Hoeyer (✉)
Centre for Medical Science and Technology Studies, Department of Public Health,
DK-1014, Copenhagen K, Denmark
e-mail: klho@sund.ku.dk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_4

every time they enter the doors to the health services. In many cases, patients also deliver tissue samples for diagnostic purposes. Simple blood samples can give rise to massive amounts of genetic sequence data and be used to identify new biomarkers, which might be of both clinical and research value. These data can be combined with other types of data such as health records, physical location, and socio-economic status extracted from non-health registers (Hood et al. 2015) and be used for planning and administration of healthcare delivery as well as research. This development is stimulated by the rapidly lowering cost of genetic sequencing technologies and electronic data storage (Richards et al. 2015; The Expert Group on Dealing with Ethical and Regulatory Challenges of International Biobank Research 2012), but also by increased demands for transparency, accountability and documentation of performance in the health services (Smith 2015). Denmark has a well-registered population and a digitalized system of performance measurement in the health services. Of potentially even greater importance for research uses of the acquired data, Denmark has a flexible legal framework where tissue samples and register data in many instances can be used for research without informed consent from the individual.

Typically, the ethical debate about health data usage focuses on the rights of individuals to determine their own participation in research. However, with this chapter I suggest that a focus on research uses and individual rights potentially misconstrues the ethical challenges involved in biomedical data usage. Data intensiveness in the health services is transforming healthcare in significant ways irrespective of research interests in data. Medical ethics therefore needs to broaden its focus. There is a need to understand why data is collected and what it is used for. And we need to move closer to actual data practices to understand what the various policy options currently discussed, such as informed consent, might produce in practice. This chapter takes some initial steps in that direction.

Many scholars embrace the notion of big data mining to capture the unfolding developments. I will suggest focusing instead on what I call ‘intensified data sourcing’. I define it as attempts at getting more data, of better quality, on more people. I chose the term ‘sourcing’ because there is more at stake than just ‘mining’ of existing data: sourcing is a dynamic process of creating, collecting, curating, and storing data while simultaneously making them available for multiple purposes, including research, governance, and economic growth. Unlike raw materials from mining, many forms of data can be retained while forwarded and thus sold or used again and again (Mayrhofer 2013). Big data debates tend to focus on research uses as a ‘revolution’ in science (Mayer-Schönberger and Cukier 2013), but the data in question is actively generated to fulfil tasks other than research.

I begin with some methodological, empirical and theoretical considerations aimed at justifying ‘intensified data sourcing’ as an alternative to the more common vocabulary of big data mining. Against this background, I outline how intensified data sourcing is unfolding in Denmark and the contestations it involves before I engage a discussion of the limitation of the on-going ethical debates about health data usage. My overall purpose is to shed light on the magnitude of, and many

competing purposes with, intensified data sourcing because I believe it is important to keep the full range of purposes with data sourcing in mind when discussing regulatory options.

2 From Data Mining to Intensified Data Sourcing

In this chapter, I take an ethnographic point of departure. It means that I base my observations on an empirical engagement with the practices that I describe (see Madden 2010). Instead of simply stating how things are, I seek to describe how I have arrived at that conviction – often by providing the anecdotal type of evidence through which we all typically learn about how others do things. I also draw on policy papers, experiences from practical work on data management committees, as well as experiences from working on data sets delivered from Statistics Denmark, and from research projects in which I have interviewed Danish researchers and patients enrolled in research because they featured in registers. My own data sourcing is as open-ended as the quantitative data practices I describe on the following pages: everything can potentially be turned into data. However, by reminding readers of particular instances through which I have learned about certain views or practices, I wish to induce a form of transparency that can sometimes be lost in big data practices where numbers come to feature as pre-existing facts.

During the past 18 months I have attended a number of meetings, conferences and workshops about 'big data'. From an ethnographic perspective one of the most striking features of these events, which all had 'big data' in the title, is how the concept attracts so many people while so few of the attendees claim to know what 'big data' is. Almost every speaker at these events denounces his or her expertise. Some point to lacking clarity about the definition of the concept and others make distanced comments to established, but vague, definitions relating to velocity, variety and volume or "data sets so large and complex that they become awkward to work with in" (each citing different sources). 'Big data' is clearly something which is happening right now in the sense that it is experienced as relating to something important: it attracts crowds. The fact that it generates so much doubt and confusion, however, makes it relevant to consider alternative conceptualizations with which to grasp the unfolding transformations. As stated in the introduction, I believe we should be talking about *intensified data sourcing*. What does this conceptualization involve and why it is relevant right now?

Rapid developments in information technologies, and lowering cost of data storage, facilitate accelerated data intensity in the *daily practices* of the health services (Buchan and Bishop 2009; Fasano 2013). Ever-more clinical practices are monitored through tracking of measurable data points, which are subsequently retained in registers or searchable files (Murdoch and Detsky 2013). Data intensity is primarily the result of activities aimed at quality assurance, performance measurement, transparency, liability protection, accounting, or to put it simply: data

sourcing is a way of running the health services (Smith 2015). Many hospitals are implementing new health informatics infrastructures to facilitate easy and pervasive re-use of medical data – primarily for administrative and clinical purposes such as quality and financial controls, documentation, communication across sectors, etc. The American company EPIC, for example, sells information platforms integrating medical records, communication between units and sectors, and accounting and documentation purposes – exactly in order to make both service delivery and secondary data usage easier.

At a recent meeting among Danish medical informatics people that I attended, the chair began the day by announcing that re-use of data beyond the care for the individual patient was of pivotal importance for the health services, and that "the important part is how we want it to influence the clinical practices through which we collect data". With collection she here refers to the formatting of the clinical record. An example of this came up subsequently in the talk of a program director working with implantation of the American EPIC platform in two Regions comprising about 40 % of the Danish hospitals.[1] The Director explained that in conjunction with EPIC they had decided to minimize text spaces in the clinical record because free text made it more difficult to use data in the daily management of the hospital. She stated that it is too difficult and time consuming to check free text when using data for administrative purposes. As I found it difficult to believe that what has typically been seen as the primary purpose of record keeping (patient care) was overruled by secondary purposes (data sourcing for administrative purposes), I wrote an email asking her for a confirmation that I had understood it correctly. As she confirmed, she added that "the decision to remove free text options is closely associated with an assessment of available tools of documentation" (email, June 8, 2015). Considering a decade of scholarly work demonstrating that from the clinical perspective free text is central to patient safety and care (Bar-Lev 2015), it is striking that from the administrative perspective templates are still seen as adequate documentation tools.

This is a good example of how data sourcing restructures care. If we used to talk about 're-use' of medical data when referring to administrative or research uses, the new healthcare informatics platforms turn the priorities around and design the record keeping instruments primarily with the 'secondary purposes' in mind. The program director also explained that EPIC was seen as a tool with which to ensure transparency and auditability down to the level of each individual physician who can get data on his or her own performance and be monitored by superiors according to the data traces they leave behind in their daily practices. The point is that data intensity is not just a new opportunity for research as if it was an add-on where we can decide *not* to use the data for research and then everything is 'back to normal'. It would be a serious misunderstanding. Key elements of care are changing irrespective of research uses, and ethics debates should therefore not be restricted to research applications but focus on how data sourcing modulates care.

[1]The two Regions are Sjælland and Hovedstaden, and further information can be found here: http://www.sundhedsplatform.dk/ (last accessed September 4, 2015).

When information platforms increase traceability of each individual patient, it becomes more relevant to also store and re-use the biological samples constantly produced in routine healthcare. Without the ability to link samples to up-to-date medical information, the samples are typically worthless. Now, ever-more tissue samples of various kinds are retained and linked to updated healthcare information (The Expert Group on Dealing with Ethical and Regulatory Challenges of International Biobank Research 2012). Besides being of value for researchers, such data practices are seen as providing new diagnostic opportunities in the daily running of the clinics (Andersen and Poulsen 2015).

Lowering cost of sequencing technologies implies that the DNA contained in the samples can be turned into data and retained in electronic files. In fact, tissue is increasingly thought of *as* data (Mayrhofer 2013). Sequence data can enter electronic repositories and be accessed from around the globe either on a pay-per-view or open-source basis (Gholami et al. 2014). In that way, intensified data sourcing operates across scales from the global to the local; from grand political and economic dimensions to the most intimate aspects of people's lives. Furthermore, the preservation of tissue and data constantly facilitate composition of new 'research populations' by combining samples from different individuals across time, geographical spaces, social classes, social networks (Holmberg et al. 2013). Data sourcing generates new potential identity markers and groups of belonging.

With all of these transformations taking place, it is no surprise that a plethora of policies are developed to govern the data and tissue flows. Specific policies tend to focus narrowly at just one of the many competing purposes, however, and they thereby neglect the complex realities of intensified data sourcing. Some policies emphasize economic interests (Organisation for Economic Co-Operation and Development 2011), others are said to protect dignity and counteract commodification of the human body (Council of Europe 1997, 2015). Some emphasize research purposes, others therapeutic or forensic purposes.[2] Some focus on data sharing (Organisation for Economic Co-Operation and Development 2007), others on protecting confidentiality and donor privacy.[3] Often the different purposes collide. Moreover, multiple agencies operating at different levels, or subdivisions of the same authorities, develop competing guidelines and rules. For example, the European Union (EU), of which Denmark is a member, seeks with its Data Protection Reform to limit data access and to protect confidentiality while promoting capitalization, while the Commission simultaneously in its role as research funder imposes data sharing and open access. Clearly the various agendas can be at odds (Kaye 2012). To understand the regulatory landscape of intensified data sourcing we therefore cannot focus on just one policy. We need to understand the everyday

[2]See, for example, the differences between the EU Tissue and Cells Directive (2004/23/EC) for biobanks aimed at therapeutic purposes, and the rules of the European Research Councils on data management plans for research biobanks, and compare to different national laws on police access to biobanks for forensic purposes.

[3]See, for example, the EU Data Protection Reform and international conventions from UNESCO, Council of Europe etc.

practices, as suggested by Mol (2002), where policies are translated into action. To understand the practices means scrutinizing the infrastructures facilitating data production and flow (Leonelli 2014). Infrastructures are never finished; they emerge through practices and never turn into fully stabilized 'things'. The point is that when I in the following begin describing intensified data sourcing in Denmark, I do not seek to account for a well-delineated system operating within a unified regulatory framework. I describe aspects of an open-ended infrastructure in-the-making that seems to be serving multiple and sometimes conflicting purposes.

3 Denmark: A Country at a Crossroad

What makes Denmark interesting for discussions about data sourcing? Long before big data became a hot topic, Denmark received attention as a country in which the entire population served as a traceable population cohort (Frank 2000). Denmark has a well-registered population where health data can be combined with data on educational and economic performance, as well as kinship and mobility patterns and many other forms of information. There are stored tissue samples from most citizens, and they can all be traced and linked to register-based information. Denmark is also (together with Singapore) the country with the highest degree of digitalization of the health services in the world (Danske Regioner 2015b, p. 9). For many years, the Nordic countries have been forerunners in biobanking and according to government estimates they are said to encompass "one fourth of the world's collective biobank capital" (Odell 2008, p. 39). With this wealth of information and traceable samples, Denmark has been tagged 'an epidemiologist's dream' (Frank 2003; Thygesen and Ersbøll 2014). As mentioned in the introduction, there are two quite specific reasons for considering Denmark an 'epidemiologists' dream': the constantly evolving *register infrastructure*, and a *lenient legal framework*. I will describe both in more detail below, and when doing so illustrate how the praised research opportunities result primarily from administrative practices. Data availability is therefore likely to persist irrespective of research uses.

3.1 The Register Infrastructure

Beginning with the register infrastructure, there is a strong tradition for register keeping. The Danish Cancer register was initiated in 1942 and claims to be the oldest of its kind in the world, but of much greater importance is the establishment in 1968 of a central register for every citizen living in Denmark called the Civil Registration System [*Centrale Personregister*] (CPR). The CPR was originally meant to serve tax purposes and it contains basic identifying information on each citizen, such as past and present addresses, family relations and marital status. These pieces of information ensure full traceability. The register thereby delivers a basic structure

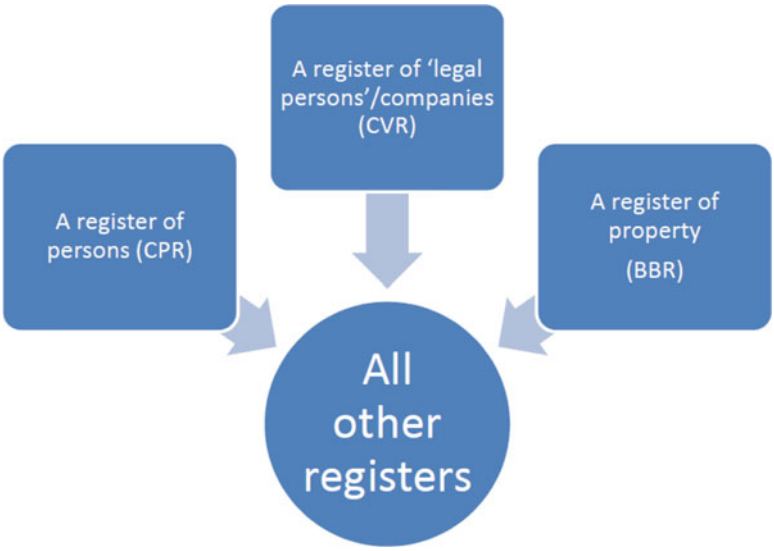

Fig. 1 The structure of Danish registers

for population traceability which since 1968 has been used in ever more additional registers containing information on people. It seems to be in constant organic growth. In addition to the register of persons (CPR), there is a register of legal entities/companies (called 'CVR'), and one of property (called 'BBR'). See Fig. 1. This basic division into owners and property reflects the taxing purposes, but as time has passed, and so many additional registers have come to use the identity numbers in these three basic registers, other purposes have become just as important for the authorities. This infrastructure today makes it possible, for example, to check if an invitation to participate in screening for cervical cancer has negative psychological implications by following the screening participants in subsequent years and see whether there are changes in their number of visits to the general practitioner, uses of psychotropics, if they encounter breaks in educational progression, increased divorce rates, etc. – to mention just one recent example of register combinations (Alexandersen 2014, p.5). With the increased digitalization of the health services and the adoption of the EPIC health informatics platform, all this information is constantly getting easier to retrieve.

3.2 The Lenient Legal System

Many countries produce more and more data and the other Nordic countries have register structures similar to the Danish system (see also Tupasela and Liede, this volume). Hence, the real competitive edge for Denmark in terms of research is the lenient legal system. Denmark has informed consent exemption rules for

biobank-based research and exemptions also from ethics approval for register-based research. Such rules are in place to minimize potential drop-out bias. It has been important to avoid drop outs in particular in relation to potentially sensitive issues. An example I have often heard Danish epidemiologists use to explain the necessity of informed consent exemptions, is that American researchers claimed that abortion caused cancer based on a study of only those women who agreed to participate in follow-up studies. When using Danish register data on all abortions and all diagnosed cancers, there was no correlation and thus it turned out that the American study was based on a selective group of women – probably those who were dependent on access to healthcare through research participation and who were exposed to other cancer risks.

A rarely mentioned reason for lenient rules on informed consent is the ubiquity of data usage in Denmark. In fact, nobody knows how often data on individual Danes are used. The authority responsible for health data delivery for research purposes, *Statens Serum Institut* (SSI), claims in a recent report written by the Danish Council of Ethics that they hand out data on practically every citizen each year, and for some patient types data are used "many times every year" (my translation, Det Etiske Råd 2015, p.9). I have also asked several people at Statistics Denmark how often a Dane can expect to supply data for statistical calculations and the guesses I have received range from ten times a year to 100,000 times.

Why is it so difficult to figure out how often data is used if everything in Denmark is so well-registered? Here we should keep in mind what I wrote about the nature of intensified data sourcing above. The different guesses probably partly reflect varying perceptions of what counts as data usage, which administrative levels they include etc. The health services have always used data, but in a paradoxical manner the increased availability of population data also seems to have made the aggregates more sensitive and more 'personal' in the public conception. In the following, I first describe some initiatives taken to promote data sourcing for research and then provide examples of public contestations and initiatives that seem to limit opportunities for data sourcing. Again, the point is to illustrate how ubiquitous data sourcing is and how research uses interact with a wider restructuring of healthcare infrastructures. I emphasize the ubiquity because I believe we need to understand the actual everyday data practices to assess the nature of the ethical problems they entail and the feasibility of the various solutions suggested.

3.3 Initiatives to Facilitate Research

The Danish Government has written into the Plan of National Growth that Danish registers and biobanks should be used to attract national and international companies (Aagaard and Lassen 2013). Data and samples cannot be sold and the authorities are not allowed to operate for profit, hence the idea is to generate economic growth by way of creating attractive environments in which companies flourish and therefore establish new places of employment. To reach the objective, work

has been initiated to create a coherent national IT infrastructure that will ease the data sourcing initiatives across sectors, institutions and domains (Danish e-Infrastructure Cooperation and Danmarks Elektroniske Fag-og Forskningsbibliotek 2015). Furthermore, targeted research service organizations have been reorganized and a National Danish Biobank has been established to facilitate easy access for national and international partners (Ministeriet for Sundhed og Forebyggelse 2013; Styrelsen for Forskning og Innovation 2013). The National Biobank is partly a physical biobank located at SSI, partly an inventory of existing biobanks located elsewhere and operated by the Regions. The Regions, who are responsible for running the Danish public hospitals, work to both construct new tissue collections and to make existing collections available for national and international public and private partners.

When research on Danish health data is discussed with policy makers at various levels, whether in closed meetings, at conferences, public hearings, or other venues, the sheer amount of available data is typically put forward as a "unique" resource giving Denmark a competitive edge. Some policymakers even talk about an "obligation" to use the data (see also Tupasela 2007, for a similar Finish example). Most researchers, however, are keenly aware that the real competitive edge stems from the lenient legal framework with its informed consent exemptions. The parliament showed its commitment to the creation of a research friendly legal framework again in 2014, when it was decided to delete an opt-out register in which approximately 16 % of the population had signed-up so that they could not be approached by researchers. It seemed surprisingly radical to just delete the register, and together with Francisca Nordfalk I decided to request access to the register and the files documenting its establishment and obliteration. We found that the register was established in 1995 in reaction to three citizens expressing concerns about surveillance, but very few used this opt out opportunity until 2001. What happened in 2001 was that a new administrative unit took over the running of the register and decided to let it feature on the forms used by citizens to inform the authorities about changes of address. On the forms it stood as an option called "Researcher protection [*Forskerbeskyttelse*]". There was no explanation of what it implied and you could register your whole family without their knowing. Interestingly, however, it was not the lacking information and autonomy associated with the registered opt out, which was discussed when Parliament deleted the register. Rather, it was the threat to research as a consequence of selection bias (Nordfalk 2015).[4] Furthermore, people in the register would still be used for register-based research irrespective of having signed up for 'researcher protection'. Registration basically meant that you could no longer be asked whether you wanted to be enrolled in research project demanding active participation.

[4]There is another opt-out register which has not been deleted yet. It is called "Tissue usage register [*Vævsanvendelsesregisteret*]" and here citizens who do not want to have their tissue used in research without consent can sign up. Less than 500 citizens have entered the register and it therefore currently does not pose a threat to population-based research.

That the authorities are well aware that consent-exemption rules are key to Danish competitiveness, has become obvious the in the negotiations of the EU Data Protection reform. Here the Danish Government has insisted on rules exempting register and biobank research from informed consent (EurActiv.com 2013). Denmark has entered into an unusual coalition with countries like Hungary in combating increased privacy rights in EU. Denmark is supportive, however, of the other central ambition with the Data Protection Reform: to attract international capital (EurActiv.com 2013; European Commission 2014). Public research funding has been used to attract one of the leading genomics companies in the world, Chinese Beijing Genomics Institute (BGI), to work on a Danish 'reference genome' (a standard genome based on which it is easier to identify variations in the Danish population). BGI has now established a European branch in Copenhagen, and I have been told by some of their collaborators that one of the reasons for their interest in Denmark is the availability of registers facilitating research that cannot be done in China.

Currently the Danish Regions and the Deans from the medical faculties of the Danish universities have joined forces to establish a large-scale whole genome sequencing project similar to the British 100,000 Genomes Project (Wynn 2014), and the population-wide genome sequencing in Iceland (Gustafsson and Farmer 2015) and on the Faroe Islands (Timmis 2011). The hope is that research and healthcare will merge and make genetic research clinically useful while simultaneously creating an important resource for research. When the policy aim is to dissolve the distinction between research and clinic, it is important that ethical analysis is not limited to research uses.

3.4 Conflicts About Data Sourcing and Initiatives to Limit Data Availability

Concomitantly with all the above mentioned initiatives to promote intensified data sourcing, the Danish authorities have, in paradoxical ways, been engaged in initiatives that seem to be aimed at imposing limits to data sourcing. Some of these initiatives relate to data security and trust issues, and some relate to changed perceptions of patient rights. I will briefly provide examples of both.

During the past decade or so, the health services have placed increasing emphasis on patient empowerment. Health services are increasingly ruled through policies formulating patient *rights*, such as a right to receive elective surgery within a particular timeframe.[5] More emphasis is given to documentation of informed

[5]Different governments have during the past 15 years introduced and modified various 'guarantees' such as entitlements to elective surgery, cancer treatment, diagnostics etc. within various time frames. Traditionally, very few health services other than abortion were presented as 'rights' – they were offered at the discretion of the treating doctor (see Hartlev 2005).

consent and avoidance of force in treatment (see also, Armstrong 2014). Patient empowerment is also sought through digital means. As mentioned above, many hospitals now use the EPIC systems where health records are to be filled in with the patient on tablets on the ward. In a new digital strategy, the Regions responsible for the public Danish hospitals aim at offering digital opportunities for patients to register the information *they* wish to share with the health services and to comment on their own health records. Already, as part of implementing patient rights through digital means, all citizens have been granted access to see what is registered in the electronic health record [*Sundhedsjournalen*] through an often-used homepage (in 2014 some 719,000 Danes looked into their own records). On this homepage, patients can also keep track of every health professional having accessed their health record. The patient role is clearly under transformation (Armstrong 2014). The point is that patient empowerment operate both through transformations of micro-practices such as changing digital interfaces and through political initiatives creating 'patient rights', and that it is deeply intertwined with the digitalization also serving to intensify data sourcing.

The potential for conflict between the two agendas, intensified data sourcing and patient empowerment, is rarely articulated. However, the increased emphasis on informed consent in relation to treatment (along with tools with which patients can monitor health professionals and data usage) makes it increasingly difficult to uphold informed consent exemptions within register and biobank research. As guardians of their own information, patients potentially acquire new forms of authority (Novas 2006). When patients do not understand who has looked into their medical records, they complain, and complaints leads to questioning of the legal framework. A very ambitious mode of data sourcing from general practice, *Datafangst* (literally 'Data catch'), was thus suspended in March 2015, when it turned out that more information was collected than what the data authorities had allowed. Four of the five Danish Regions announced in January 2015 that now also non-clinical uses of hospital records would be suspended until it was determined whether uses without informed consent were legal (Hildebrandt 2015). While few Danes have heard about the suspension of data sourcing from hospital records, the case from general practice has been widely publicized (Stræde 2014). According to a study of social media reactions to the case (Westergaard and Skovgaard 2015), opposition to *Datafangst* has been promoted by general practitioners contesting administrative uses of data to monitor their performance. The GPs have not only written numerous opinion pieces, they have also set up a new association aimed at giving patients control of their own data (see http://patientdataforeningen.dk/) along with several initiatives on social media, which have generated thousands of likes and comments from citizens expressing discomfort with uses of their health data (Westergaard and Skovgaard 2015).

If we consider what features in the news, it is probably reasonable to assume that the public discomfort with data sourcing, which is recorded in the study by Westergaard and Skovgaard and in my own interviews with patients enrolled in research (see below), is fuelled also by numerous recent examples of data leakages, instances of hacking, and anxieties related to surveillance. There are elements of

the specific Danish context which are important to notice when we discuss data security and leakages. The running of the CPR organization has been outsourced to a private company, CSC, which has then been sold to an equity fund. The number of employees has been drastically reduced and in the course of the many changes, CSC has had some unfortunate problems with data security. Another case of data leakage receiving massive media attention was the revelation that the Danish gossip magazine, *Se og Hør*, illegally bought information about the royal family and other notabilities from a person having access to their credit card information from a nationally initiated credit card company, *Dankort*, which has also been outsourced. Another case receiving public attention was when a private company selling internet security options did a little investigation for the national broadcasting company which revealed 1492 CPR numbers illegally featuring on the homepages of 21 Danish municipalities (Bengtsson 2014). The report indicated that the sensitive information typically featured in relation to publication of citizen complaints, which implies that it is the transparency and empowerment agenda that here leads to confidentiality breaches.

In September 2014, I interviewed 21 persons who were identified through registers and invited to participate in genetic research. They were keenly aware of media stories about data leakages. Importantly, however, they rarely distinguished between various agencies or organizations. They would mention the cases of Se og Hør, CSC and the politically sensitive issues of national security surveillance (e.g. by the US National Security Agency) and attempts of revealing power through data leakages (e.g. WikiLeaks). Such non-health issues may thus potentially intervene in public perceptions of health registers and electronic patient records and shape the direction of intensified data sourcing in the biomedical arena. Controversies with no official relation to health data might gradually influence public perceptions of control and comfort in relation to health data management. Still, each of the interviewed participants had decided to trust the researchers inviting them. Several of them remarked that data might leak, but they assumed nobody would care to pry into their medical information: "I just don't think I'm interesting enough for anyone to go through that trouble", as one of them said. Hence, rather than expressing trust as a form of personal confidence, they tend to express a form of fatalistic choice ("go as it may") more in line with Luhmann's classic conception of trust as mechanism to reduce complexity in modern society (Luhmann 1999). I pose this remark on trust because I, throughout the past 15 years, again and again have heard medical researchers, politicians and administrators claim that "Danes" trust them. I am not claiming a sudden crisis of trust. Indeed we cannot know to what extent there ever was trust among the Danes. Nor can we say that there is no trust in the current system. Patients still participate in research, they go to their doctors, hospitals continue to keep records, everything continues to work more or less as it used to. Nevertheless, something is changing. Denmark seems to have arrived at a crossroad with no clear path ahead and no clear understanding of how citizens should manoeuvre between being patients and data subjects. In the following I turn to the ethical debates emerging at this crossroad. I present some of the 'solutions' that policymakers consider, and discuss their merits in lights of the practices just described.

4 Ethical Debates and Their Limitations

It is striking how the current intensification of data sourcing happens in parallel with a sort of public 'ethical awakening' with regards to data. Though I have shown how health data usage is not new at all, biomedical data was not previously seen or discussed in Denmark as problematic. In fact, the very conception of data seems to be changing. Previously medical data was rarely discussed as information about individual citizens. It was data about the work and performance of health professionals, data for billing purposes, data on quality etc. The data record happened to contain information about individuals, but the information about specific persons was rarely articulated as the defining characteristic. With the new form of ethical debate about health data, however, our conception of the health data as such seems to be changing.

This awakening takes place at both international and national levels. It comes across in the literature, with volumes like this one, and the report on health data from the Nuffield Council (Richards et al. 2015). It also comes across in policy making, such as the new declaration on health data from the World Medical Association, and other initiatives discussing data security and the rights and integrity of data subjects (e.g. the EU Data Protection Reform). At the national level, the Danish Council of Ethics recently issued a report on health data (Det Etiske Råd 2015), and the Danish Regions issued both a data management policy and a policy for Data usage in 2015 (Danske Regioner 2015a, b). The new data awareness also comes across in the social media activism mentioned above. Finally, it comes across in an intense organizational activity aimed at reorganization of data practices – which are now often discussed as the making of 'ethics policies'. I personally experience the organizational demands as a member of several administratively appointed committees working in parallel to enhance data management practices at various levels at my home university.

There are several things worth noticing about this process of making ethics policies for data. First, organizational actors at the national level tend to suggest "solutions" aimed at the *individual*, though discussing data sourcing that operates at the population level. Second, the ethical attention seems to focus on *research* uses though the data in question is mostly generated for purposes other than research and therefore most likely will continue to exist and modulate care irrespective of the rules set for research. In the following I will discuss the problems with first the focus on the individual and then the one on research uses.

4.1 The Individual and the Population

With solutions aimed at the individual I think, of course, of suggestions of introducing informed consent with respect to data usage. Informed consent is primarily designed to let an individual assess the risks associated with participating in, for example, clinical trials (Faden and Beauchamp 1986). It has been extended

to ever more practices and is today a cornerstone also of the movement towards patient empowerment in everyday clinical practice. But who can be expected to provide informed consent to data usage once a year, ten times a year, or 100,000 times a year (depending on which of the figures above we should trust)? Even if the figure could be brought down from 100,000 to something like ten times by limiting what is included when counting (e.g. by only counting specific forms of data usage for particular forms of research), it will be a substantial amount of consent procedures that individuals will have to undergo for today's practices to continue. The Danish experiences with 'opt out' (the register which was recently deleted) do not point to a successful alternative to informed consent, but just as importantly those experiences indicate what happens when opting in or out becomes a matter of ticking boxes while you are in the middle of doing something else. We cannot expect individuals to mind the problems of the population in a manner that will make them consider research participation in any depth when it happens so often. Either we need to accept that the current research activity cannot continue or we need to install alternative procedures to safeguard the interests of the contributing individuals. Informed consent will not work. The Danish Council of Ethics suggests in their recent report to consider a 'meta-consent' (with traits of what Jane Kaye et al 2015, has developed as dynamic consent) where individuals can indicate their consent preferences online – but the ethical attention in this way remains focused on what the individual decides, not what the individual has at stake.

Already, informed consent has been intensely criticized on numerous accounts. In relation to biobanking, it is impossible to 'inform' about unanticipated future technological opportunities (Arnason 2004; Hansson et al. 2006; Lipworth et al. 2006; Maschke 2006; Nõmper 2005; Steinsbekk et al. 2013; Wendler 2006). Informed consent has also been seen as inadequate for genetic biobanking when consent is given by individuals, while the samples reveal information about families and groups (Beskow et al. 2001). Furthermore, it has long been known, that patients and research participants rarely understand, remember or use the information provided (Hoeyer and Hogle 2014; Lidz et al. 1985). Informed consent is therefore accused of deflecting attention from more important moral issues and reducing ethical reflection to procedural collection of signatures (Brekke and Sirnes 2006; Corrigan 2002, 2003).

What is informed consent then supposed to do? Traditionally informed consent is expected to protect the individual and preserve the individual's autonomy (Faden and Beauchamp 1986). In relation to *protection*, the Danish Council of Ethics discusses both protection against informational risks, in the sense of harm encountered as a consequence of data leakage (e.g. limited access to insurance) and in the sense of breaches of privacy, by which they think of emotional harm disturbing the well-being of the individual. With *respect for autonomy*, the purpose is to let each individual influence what they support. At various meetings, I have heard several times an analogy to charity of the type: "You have a right to decide what your money is used for and whether they should go to Save the Children or Red Cross. The same ought to apply to your tissue or your data." Actually, it is not absolutely true about money for aid work if we consider the fact that

Danish development aid is primarily derived from taxation, but leaving that aside my point here is that it is problematic when respect for autonomy is conflated with the objective of protection. In fact, the introduction of individual informed consent demands to enhance autonomy might very well undermine the protection against data leakage because systems will need to uphold a strong data linkage to each individual – without data linkage, it is impossible to give individual consent. In Denmark it has been suggested using the site www.sundhed.dk where each individual can log in and see their own health records to allow individuals also to administer consent for research uses. However, the more an individual can access, the more entry points will be generated. Such entry points can then be used also by hackers and others looking for access to data. The online society where everybody can access everything about themselves is also online 24/7 for the uninvited guests and intruders. Conversely, when anonymization is suggested as an alternative to informed consent it might protect them, but it will hardly serve to protect their autonomy (Mittelstadt and Floridi 2016). We might therefore need to think more carefully about the differences between procedures enhancing autonomy and those ensuring protection, and articulate more clearly the ambitions with the policies adopted. Informed consent cannot take care of it all and irrespective of the solutions policymakers adopt; they all come at a cost.

While the focus in ethics debates on informed consent has tended to narrow down the discussion to autonomy and protection, Mittelstadt and Floridi (2016) argue that data practices might have implications for patients in other ways. When population data are used to generate guidelines for future prevention or treatment plans, the data have looping effects for the people delivering data and samples (Bauer 2014; Hacking 1995; Holmberg et al. 2013). New diagnostic information might occur, and new risk profiles emerge. Profiling through big data is often aimed at prevention. Individuals become subjected to new forms of advice and have to do things differently – rarely because the individual will fare better, but because 25 out of 1000 will. Prevention typically has effects at the level of the population. Furthermore, it is necessary to acknowledge that sometimes public health advice aimed at prevention turns out to be all wrong. There is a long and unfortunate history of preventive campaigns that turned out to build on inadequate evidence. We need to be aware that this will also be the case for some of the 'big data insights' that currently promises to deliver personalized health optimization (Hood and Flores 2012). Even when the advice is medically sane, some will object to health promotion as a lifelong preoccupation (Rose 2007). Not everyone wants research to enter the clinic. Still, in the current environment the ambition is to dissolve the distance between 'bench [or computer] and bedside' and make data usage part and parcel of clinical practice down to the level of the individual patient. Therefore, we need to keep asking in relation to the various initiatives: For whose sake?

Deleuze viewed the turn to life-long prevention as a special form of control (Deleuze 1990) and Foucault has inspired a range of studies focusing on the medical gaze as involving biopolitical surveillance and subjectification (Foucault 2000). Nevertheless, it is important not to over-generalize research tendencies. There are many competing research agendas using the same resources and it is never obvious

which agenda that will have most influence (Leonelli 2014). There is a need to explore what data sourcing eventually produces, for whom, and why. When doing this, it is again important to remember that data are mostly construed and used for purposes more mundane than research: they are indeed supposed to serve purposes of control, but in a somewhat less encompassing sense than Foucault and Deleuze suggested, namely to document the quality of care. In many cases, it might very well be in the best interest of the individual patient. And finally, we need to accept that it is not only patients who hold legitimate interests in storage of data. Health professionals need to document their practice, both to learn and to protect themselves against potential accusations from discontent patients.

4.2 Why Focus on Research Uses of Data?

In its recent report, the Danish Council of Ethics decided to discuss only research uses of health data and suggested considering various consent options of research uses of data. Even with informed consent (or opt out) for research uses, however, other forms of data usage will continue – and therefore we must question whether a demand for informed consent for research would really make any difference for the individual in terms of the informational risks associated with data usage? Most researchers get only anonymized data sets, hence the number of people having access to identifiable data will not necessarily decrease though research opportunities do. The most important problem with the current focus on research uses of data, however, relates to the way in which it imposes a certain form of blindness to the more pertinent risks for the individual associated with intensified data sourcing. When focusing ethical debates on research uses, the real changes to healthcare installed through intensified data sourcing remains largely unaddressed. For example, the fact that EPIC platform reduces free text areas in the health record keeping to support data surveillance for administrative purposes, will have significant implications for the care of the individual. It has nothing to do with research uses, but it introduces changed priorities in record keeping whereby 'secondary uses' comes to define the technical options available for the 'primary uses' of data. It is a fundamental reconceptualization of the purpose with record keeping which ought to be discussed. As long as ethical debates focus on research uses, it is not.

The World Medical Association suggests informed consent for *all* forms of data usage and does not limit itself to research uses. Imagine that this would actually be the case – each individual patient could decide that biomedical data cannot be stored without their consent – what would the consequences be for that individual? There would be clinical consequences, as the individual having opted out of record keeping would have to keep personal track of specialized medical information in order to get proper medication etc. In emergencies, or if the patient is unconscious or loses mental abilities, treatment opportunities could be severely hampered because treatment in these cases relies on access to the previously recorded medical history.

If patients have opted for storing less information about themselves, will they then also lose entitlements related to complaints and medical liability? Is it fair to expect health professionals to accept complaints if they are not allowed to retain data on treatment? There is no risk-free form of healthcare.

5 Concluding Remarks

With this chapter I have argued for the need to acknowledge the wider context for big data debates by way at looking at the practices through which data are created and used. Because data are not just 'mined' but actively constructed, and because they serve many competing purposes among which research is often just a minor element, I have suggested conceptualizing the ongoing transformations as *intensified data sourcing*. My central point has been that when we look beyond research uses, we might better understand why data sourcing takes place and thereby comprehend the nature of both the problems and benefits data sourcing involves for the people delivering the data.

I have presented a national case study from a country which in many ways has tried to take the lead in intensified data sourcing. From an outside perspective, Denmark can appear as a research radical country because many initiatives have been taken to promote research, including the abolition of an opt-out register. It is important, however, to note that the overall aim of enhancing data availability relates to purposes other than research. The intensive mode of data sourcing has moved the country to a point where some initiatives, such as *Datafangst*, have been found unlawful. Data leakages and hacking mostly unrelated to medical research have subjected data flows to heated public debate and scrutiny and brought Denmark to a crossroad where the public legitimacy of data usage cannot be taken for granted anymore.

Intensified data sourcing is ubiquitous in Denmark. Nevertheless, several ethics policies, including the recent report from the ethics council, focus on *research* uses and suggest remedies to potential problems that focus on the *individual*. Informed consent is mentioned again and again as a policy 'solution' to all types of problems (Hoeyer 2009). Individual informed consent, however, is not compatible with the actual data practices. Today, the health services are managed through data sourcing, and data does not disappear just because research uses are limited. Even if consent procedures are limited to *research* uses of data, the amount of consent procedures each individual should engage in will be so significant that we cannot expect people will pay due attention to the actual information provided. Furthermore, there is a need to reconsider the double role typically attributed informed consent of *protection* and a means for respecting *autonomy* (Laurie and Hunter 2013, p.38). Some solutions aimed at enhancing autonomy, might work counter to the aim of protection, and some measures aimed at protection (such as increased anonymization) will limit the ability of each individual to exert autonomy.

At a more basic level, however, we should initiate a discussion about the nature of the personal stakes in biomedical data practices. How and why did we come to re-interpret data in the health services as *personal* data? We seem to be in a process of negotiating not only rights and duties of citizens, but also what a person is and where we encounter persons. The medical record is no longer an object of internal clinical practices related to auditing, quality assurance, management and research; the virtual cloud of data seems to gradually become personified as yet another representation of a named person. It is of course this insistence on the image of a 'person' in the data cloud that makes informed consent appealing as a passage point from virtual cloud to research population. But can we really expect actual people of blood and flesh to spend the time needed to assess each research project – or all the other uses of health data? What are the consequences for the individual, if opting out becomes an easy solution, an online click and you are left in peace? Will individuals end up losing rights if they have opted for 'no data trail' at some earlier point? I believe we need to think through carefully the consequences of providing individuals the right to delete data otherwise retained in the health services and do so in light of a conceptually nuanced consideration of the nature of the retained data.

These challenges are not just Danish, but Denmark is at the forefront of experiencing them. In its work with data protection, the European Union will need to take into account the very different forms of data sourcing that are currently developing in different European countries, and therefore national case studies such as this one are needed. It is through engagement with concrete data practices we learn what is at stake for the people involved, and a proper ethical analysis should take these stakes into account.

References

Aagaard, J., and L.H. Lassen. 2013. *Dansk databank skal tiltrække udenlandsk sundheds-forskning*. Business.dk, http://www.business.dk/oekonomi/dansk-databank-skal-tiltraekke-udenlandsk-sundhedsforskning

Alexandersen, H. 2014. Forskningsservice gennem 25 år, ed. L. Thygesen, I. Thaulow, and C. Zangenberg. København: Danmarks Statistik.

Andersen, T.B., and M.-B.J. Poulsen. 2015. *Handlingsplan for Projekt Personlig Medicin*, 1–9. Copenhagen: Danske Regioner.

Armstrong, D. 2014. Actors, patients and agency: A recent history. *Sociology of Health & Illness* 36(2): 163–174.

Arnason, V. 2004. Coding and consent: Moral challenges of the database project in Iceland. *Bioethics* 18(1): 27–49.

Bar-Lev, S. 2015. The politics of healthcare informatics: Knowledge management using an electronic medical record system. *Sociology of Health & Illness* 37(3): 404–421.

Bauer, S. 2014. From administrative infrastructure to biomedical resource: Danish population registries, the "scandinavian laboratory," and the "epidemiologist's dream". *Science in Context* 27(Special issue 02): 187–213.

Bengtsson, K.L. 2014. *Pressemeddelelse*. Copenhagen: Udbudsvagten.

Beskow, L.M., W. Burke, J.F. Merz, P.A. Barr, S. Terry, V.B. Penchazadeh, L.O. Gostin, M. Gwinn, and M. Khoury. 2001. Informed consent for population-based research involving genetics. *JAMA* 286(18): 2315–2321.

Brekke, O.A., and T. Sirnes. 2006. Population biobanks: The ethical gravity of informed consent. *BioSocieties* 1: 385–398.

Buchan, I., and J.W.C. Bishop. 2009. A unified modeling approach to data-intensive healthcare. In *The fourth paradigm*, eds. T. Hey, S. Tansley, and K. Tolle, 91–97. Redmond: Microsoft Research.

Corrigan, O. 2002. *Trial and error: A sociology of bioethics and clinical drug trials*. London: University College of London.

Corrigan, O. 2003. Empty ethics: The problem with informed consent. *The Sociology of Health and Illness* 25(3): 768–792.

Council of Europe. 1997. *Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention of human rights and biomedicine*. Strasbourg: Council of Europe.

Council of Europe. 2015. *Council of Europe convention against trafficking in human organs*. Santiago de Compostela: Council of Europe.

Danish e-Infrastructure Cooperation, and Danmarks Elektroniske Fag-og Forskningsbibliotek. 2015. *National strategi for forskningsdata management 2015-2018*. København. Lyngby: DelC.

Deleuze, G. 1990. Kontrol og tilblivelse. In *Forhandlinger 1972-1990*, ed. G. Deleuze, 203–220. Frederiksberg: Det lille forlag.

Det Etiske Råd. 2015. *Forskning i sundhedsdata og biologisk materiale i Danmark*. København: Det Etiske Råd.

EurActiv.com. 2013. *Data protection reform in peril as Germany stymies deal*. http://www.euractiv.com

European Commission. 2014. *Progress on EU data protection reform now irreversible following European Parliament vote*, 1–10. Brussels: European Commission.

Faden, R., and T.L. Beauchamp. 1986. *A history and theory of informed consent*. Oxford: Oxford University Press.

Fasano, P. 2013. *Transforming health care: The financial impact of technology, electronic tools and data mining*. Oxford: Wiley.

Foucault, M. 2000. *Klinikkens Fødsel [The birth of the clinic]*. Copenhagen: Hans Reitzels Forlag.

Frank, L. 2000. When an entire Country is a cohort. *Science* 287(5462): 2398–2399.

Frank, L. 2003. The epidemiologist's dream: Denmark. *Science* 301(5630): 163.

Gholami, A., A.-S. Lind, J. Reichel, J.-E. Litton, A. Edlund, and E. Laure. 2014. Privacy threat modeling for emerging BiobankClouds. *Procedia Computer Science* 37: 489–496.

Gustafsson, J., and E. Farmer. 2015. *The genomic portrait of a nation*. Delaware: Amgen.

Hacking, I. 1995. The looping effects of human kinds. In *Causal cognition: A multidisciplinary debate*, ed. D. Sperber, D. Premack, and A.J. Remack, 351–394. Oxford: Clarendon.

Hansson, M.G., J. Dillner, C.R. Bartram, J.A. Carlson, and G. Helgesson. 2006. Should doners be allowed to give broad consent to future biobank research? *The Lancet Oncology* 7: 266–269.

Hartlev, M. 2005. *Fortrolighed i Sundhedsretten – et Patientretligt Perspektiv*. København: Forlaget Thomson A/S.

Hildebrandt, S. 2015. Ny lovfortolkning er en bombe under kvalitetsarbejdet. *Dagens Medicin* 9: 6–7, April 10, 2015. Copenhagen.

Hoeyer, K. 2009. Informed consent: The making of a ubiquitous rule in medical practice. *Organization* 16(2): 267–288.

Hoeyer, K., and L.F. Hogle. 2014. Informed consent: The politics of intent and practice in medical research ethics. *Annual Review of Anthropology* 43: 347–362.

Holmberg, C., C. Bischof, and S. Bauer. 2013. Making predictions: Computing populations. *Science, Technology & Human Values* 38(3): 398–420.

Hood, L., and M. Flores. 2012. A personal view on systems medicine and the emergence of proactive P4 medicine: Predictive, preventive, personalized and participatory. *New Biotechnology* 29(6): 613–624.

Hood, L., J.C. Lovejoy, and N.D. Price. 2015. Integrating big data and actionable health coaching to optimize wellness. *BMC Medicine* 13(4): 1–4.

Kaye, J. 2012. The tension between data sharing and the protection of privacy in genomics research. *Annual Reviews of Genomics and Human Genetics* 13: 415–431.
Kaye, J., E.A. Whitley, D. Lund, M. Morrison, H. Teare, and K. Melham. 2015. Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics* 23(2): 141–146.
Laurie, G., and K. Hunter. 2013. *Guthrie cards in Scotland: Ethical, legal and social issues*. Edinburgh: The Scottish Government.
Leonelli, S. 2014. What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society* 1(1): 1–11.
Lidz, C.W., A. Meisel, and M. Munetz. 1985. Chronic disease: The sick role and informed consent. *Culture, Medicine and Psychiatry* 9(3): 241–255.
Lipworth, W., R. Ankeny, and I. Kerridge. 2006. Consent in crisis: The need to reconceptualize consent to tissue banking research. *Internal Medicine Journal* 36: 124–128.
Luhmann, N. 1999. *Tillid – en mekanisme til reduktion af social kompleksitet*. København: Hans Reitzels Forlag.
Madden, R. 2010. *Being ethnographic. A guide to the theory and practice of ethnography*. London: Sage.
Maschke, K.J. 2006. Alternative consent approaches for biobank research. *The Lancet Oncology* 7: 193–194.
Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A revolution that will transform how we live, work and think*. London: John Murray.
Mayrhofer, M. Th. 2013. About the new significance and the contingent meaning of biological material and data in biobanks. *History and Philosophy of the Life Sciences* 35(3): 449–467.
Ministeriet for Sundhed og Forebyggelse. 2013. *STARS* – Strategisk Alliance for Register- og Sundhedsdata*, 1–2. Copenhagen: Ministeriet for Sundhed og Forebyggelse.
Mittelstadt, B.D., and L. Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.
Mol, A. 2002. *The body multiple: Ontology in medical practice*. London: Duke University Press.
Murdoch, T.B., and A.S. Detsky. 2013. The inevitable application of big data to health care. *JAMA* 309(13): 1351–1352.
Nõmper, A. 2005. *Open consent – A new form of informed consent for population genetic databases*, 5–260. Tartu: Tartu University.
Nordfalk, F. 2015. *Forskerbeskyttelsen i Danmark 1995-2014*, 1–79. Copenhagen: Københavns Universitet.
Novas, C. 2006. The political economy of hope: Patients' organizations, science and biovalue. *BioSocieties* 1(3): 289–305.
Odell, M. 2008. *Ett lyft för forskning och innovation*. Stockholm: Regeringen.
Organisation for Economic Co-Operation and Development (OECD). 2011. *The bioeconomy to 2030: Designing a policy agenda*. Paris: OECD.
Organisation for Economic Co-Operation and Development (OECD). 2011. *The bioeconomy to 2030: Designing a policy agenda*. Paris: OECD.
Regioner, Danske. 2015a. *Regionernes politiske linje for informationssikkerhed*. Copenhagen: Danske Regioner.
Regioner, Danske. 2015b. *Sundhedsdata i spil*. Copenhagen: Danske Regioner.
Richards, M., R. Anderson, S. Hinde, J. Kaye, A. Lucassen, P. Matthews, M. Parker, M. Shotter, G. Watts, S. Wallace, and J. Wise. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*. London: Nuffield Council on Bioethics.
Rose, N. 2007. *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century*. Princeton: Princeton University Press.
Smith, P.C. 2015. Reflecting on 'Analytical perspectives on performance-based management: An outline of theoretical assumptions in the existing literature'. *Health Economics, Policy, and Law* 10(4): 479–483.

Steinsbekk, K.S., B.K. Myskja, and B. Solberg. 2013. Broad consent *versus* dynamic consent in biobank research: Is passive participation an ethical problem? *European Journal of Human Genetics* 21: 897–902.

Stræde, M.K. 2014. Oplysninger om patientdatabase belaster myndigheder. *Information*: 8–9, November 19, 2014. Copenhagen.

Styrelsen for Forskning og Innovation. 2013. *Registerforskning – nye muligheder og nye udfordringer*, 1–30. Copenhagen: Styrelsen for Forskning og Innovation.

The Expert Group on Dealing with Ethical and Regulatory Challenges of International Biobank Research. 2012. *Biobanks for Europe: A challenge for governance*. Brussels: European Commission.

Thygesen, L.C., and A.K. Ersbøll. 2014. When the entire population is the sample: Strengths and limitations in register-based epidemiology. *European Journal of Epidemiology* 29(8): 551–558.

Timmis, O. 2011. Faroe Island to be the first to sequence an entire nation. *Bionews*. vol 628, London.

Tupasela, A. 2007. Re-examining medical modernization – Framing the public in Finnish biomedical research policy. *Public Understanding of Science* 16(1): 63–78.

Wendler, D. 2006. One time general consent for research on biological samples. *British Medical Journal* 332: 544–547.

Westergaard, A.W., and L.L. Skovgaard. 2015. *Databasen der delte vandene: Værdirationaler i debatten omkring Dansk AlmenMedicinsk database*, 1–51. Copenhagen:Københavns Universitet.

Wynn, S. 2014. 100,000 Genomes: Impacting real lives. *Bionews* 780. London.

A Critical Examination of Policy-Developments in Information Governance and the Biosciences

Edward Hockings

Abstract This chapter provides a contextualisation of policy developments in the biosciences, health-research and information governance in relation to societal tendencies. The initiatives considered include the Clinical Research Practice Datalink, the Health and Social Care Information Centre, the 100,000 Genome Project, the introduction of personalised medicine, and the relaxation of the information governance regulatory regime. It is argued that we are witnessing a shift from rights-based approach to the adjudication of competing claims, in which benefits to the economy, for example, are seen as goods to be balanced with a data subject's right to privacy and confidentiality. The greater weight that economic interests now possess in information governance and the biosciences has substantive implications for future policy in this area. This, along with the new powers of access by the Government, moves us into unexplored terrain and generates novel ethical challenges. Though, this chapter advocates an approach to policy and governance that proceeds along deliberative and democratic lines, the developments of concern to this chapter are evidence of the challenges that lie ahead.

1 Introduction

The ambition by the Government to turn the UK into a leader in the biosciences has led to major infrastructural developments, creating unprecedented access to patients' medical information held across the NHS, and to whole-sequenced genomic data collected from NHS users. This has been accompanied by the relaxation of the governance of medical data and the introduction of permissive normative guidelines for the use of biomedical data. Forming part of an innovative economic and health policy, the introduction of personalised medicine into NHS healthcare from 2017, the 100,000 Genome Project and other data intensive initiatives, will add value to the bioscience and health-research sectors, lead to better management of NHS resources and an improvement in the quality of healthcare, and contribute to the maturation of genomics. However, policy developments have tended to narrow

E. Hockings (✉)
University of the West of Scotland, High Street, Paisley PA1 2BE, UK
e-mail: Edward.Hockings@uws.ac.uk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_5

down decisions, focusing on such things as delivering economic growth, rather than taking broader concerns into account. Commercialisation now carries incomparably greater force in the governance of medical and biomedical data. This is tied to a more general trend in which the economic paradigm has come to dominate policy relevant to the biosciences in the UK. Short-term economic motives, however, are likely to have long-term ethical and socio-legal consequences. This chapter takes a holistic view of these developments, focusing in particular on the novel challenges engendered by the relaxation of the regulatory regime, and the commercialisation of biomedical 'Big data.'

We are witnessing a shift in the governance of medical and biomedical data, from a rights-based approach to the adjudication of competing claims, in which benefits to the economy, for example, are seen as goods to be balanced with a data subject's right to privacy and confidentiality. These unprecedented levels of access by Government and private sector, give rise to new powers and new responsibilities. As these initiatives have been driven by the government and the private sector, we cannot expect these new powers to be used in ways which reflect the interests of society as a whole, but rather, sectional interests and those of Government. The embedding of commercialisation in the governance of medical and biomedical data is likely to lead to the widening of the uses of whole-sequenced genomes for commercial purposes. Further, hitherto unseen levels of access by the Government might also result in an expansion of the ends for which biomedical data is used. The view is advanced that the relaxation of the governance model must be supplemented with adequate governance arrangements to oversee the uses of this information, by whom and for what reasons.

Forces on the ground, such as the public relations strategy employed by successive governments', have meant that the conditions for public deliberation about the nature, risks and wider implications of the 100,000 Genome Project, and related initiatives, have not been in place. Further, the deliberate obfuscation of the associated risks and the wider trajectory of policy have facilitated the locking in of commercialisation in the biosciences. Redressing the democratic deficit that is already in place requires, as a first step, acknowledging the salience of the socio-economic context of modern societies in shaping governance models and policy decisions. The stipulation that bioscience policy and governance must be deliberative and democratic (Nuffield 2012) should be situated within a macro-level view of society. In particular, the tendency towards de-democratisation (Habermas 2015) and the expansion of the reach of market mechanisms and market thinking (Sandel 2012). Novel developments such as the advances in the biosciences are, in great part, being driven by market thinking and concomitantly, sectional interests are being privileged over democratic decision and deliberative engagement with the wider trajectory of bioscience policy. As the balance between the market and democracy is out of synch (Habermas 2015), modern societies are failing to engage with issues of paramount importance (Sandel 2012).

This, then, raises some crucial questions. As the biosciences could result in harms at a public scale (Nuffield 2012), there is a significant public interest in these. Is there

moral warrant to afford priority to democratic decision and deliberative engagement over sectional interests and market thinking? Further, are modern societies capable of a practical discourse in regard to regulatory issues in the biosciences and the trajectory of bioscience policy? Lastly, if modern societies are capable of consensus formation, on what criteria might such a process be considered morally binding? These avenues of enquiry will be explored through Jürgen Habermas' theory of communicative action and his discourse theory of ethics, which, it will be argued, provide normative and theoretical grounds from which the policy and regulatory developments of concern to this chapter will be shown to be ethically problematic.

Placing particular emphasis on the driving forces behind them, this chapter begins with a detailed account of policy developments. This is followed by an assessment of the changes to the landscape of information governance, in particular, the newfound weight that commercialisation carries therein. The following section considers the impact of the biosciences on values, and charts a shift from a sanguine acceptance about attendant changes, to a more cautious approach, which is sensitive to the social, economic and legal context in which the biosciences are emerging and the implications of their introduction into society. This, furthermore, has been accompanied by the view that the public must be given a central role in the development of bioscience policy and regulation. Consideration is then given to the features such an approach ought to possess. The need for reflexivity, openness and inclusiveness is then contrasted to the way policy and regulatory developments have unfolded. Specific focus is placed on the public relations strategy that was employed; the deliberate obfuscation of the risks, an absence of meaningful public discourse, and democratic deficit at the heart of these developments are central to the critical analyses. The proceeding section anchors an inclusive, open and reflective approach to bioscience policy in Habermas' theory of communicative action and his discourse ethics.

This leads us into to a detailed explication of the risks, which includes a critique of the narrowly focused discussion of these in the relevant literature. Consideration will then be given to the unprecedented responsibilities for Government and the new opportunities for commercialisation that is engendered by the relaxation of the governance of sensitive information.

2 Policy Developments: An Overview

At the turn of this Century, a memorandum by Glaxo Wellcome, submitted to the Select Committee on Science and Technology, House of Lords (2000), advanced the view that "the UK has an opportunity through the National Health Service (NHS) system of 'tracking' patients and using electronic medical records to establish a valuable genetic research database." At the same time, written evidence by SmithKline Beecham to the Select Committee on Health, House of Commons (2000) stated that the "NHS represents a singular but under-utilised resource

for population genetics, and healthcare informatics more generally". The same submission stated that what was "now required is the political will to tackle the issue of public acceptability." In 2007, evidence submitted by the Association of the British Pharmaceutical Industry to the Select Committee on Health, House of Commons (2007), stated that there is an "international race for benefit and competitive advantage in research where the UK could have a significant Unique Selling Point (USP), if research interests are given priority . . . making use of the full electronic patient record will provide substantial benefits to patients, the NHS and the economy".

By 2008, as part of the 'NHS National IT Programme', electronic medical records (Summary Care Records) were rolled-out across England and Wales. Through the 'Secondary Uses Service', every NHS user, by default, become a research subject. With a change of Government, the NHS National IT Programme was discontinued; however, the ambition, to quote from the Plan for Growth 2012 (PfG), Department for Business, Innovation and Skills (BIS), of "using e-health record data to create a unique position for the UK in health research", remained unchanged. The 2011 Strategy for UK Life Sciences (SUKLS), Department for Business, Innovation and Skills (BIS), set forth the vision for the exploitation of the NHS as a data source and, as part of the strategy, committed to consult on an amendment to the NHS Constitution which would see the introduction of a default assumption that "data collected as part of NHS care can be used for approved research, with appropriate protection for patient confidentiality" (2007, p.3)". The Government Plan for a Secure Data Service 2012 affirmed that, as the "NHS could offer unique opportunities for this country's international competitiveness in health research," it would "create the capacity to draw on the power of large linked data sets on a scale unprecedented here or elsewhere in the world."

Beginning in 2012, the Conservative Government launched three major initiatives. The Clinical Practice Research Datalink (CPRD) was established, which provides NHS clinical data to the research and life sciences communities. Collecting comprehensive information about treatments and care from across health and social care, which feed into clinical auditing and the planning of new health services, and, also, it creates custom data sets – about which more will be said later – the Health and Social Care Information Centre (HSCIC) was set up, in 2013. Contemporaneously, a report by the Department of Health (DoH) – Building on our Inheritance: Genomic Technology in Healthcare 2012, called for the NHS to consider significant modernisation and development of genetic testing services. Months later, David Cameron PM, announced that a paradigm shift in healthcare, which would see genomics used across the NHS, making the UK a global leader in genomics, enhancing clinical care and contributing to economic growth – will take place. A Freedom of Information disclosure by the Department of Health confirmed the Government's intentions; as the "UK is well placed to play a world-leading role in this next phase of the biomedical revolution . . . a central repository for storing genomic and genetic data and relevant phenotypic data from patients" – which would support the growth of UK genomics and bioinformatics companies" – would be created.

Later that year, the government went public about the detail; focussing on patients with inherited/rare diseases and cancer, 100,000 genomes would be sequenced by 2017 (Walker 2012). In 2013, 'Genomics England', a limited company responsible for the sequencing of the personal genomes and the creation of a dataset of "whole genome sequences, matched with clinical data", at "a scale unique in the world", was created (Genomics England 2013). Shortly after, eleven Genomic Medicine Centres, which would provide engagement and feedback to those participating in the 100 K GP, were rolled-out across England. Creating the capacity to link phenotypic and clinical data with whole-sequenced genomes, are the first steps in the transition to a model of healthcare based on prediction, prevention and personalisation, which would ultimately see, in the words of the Health Secretary, the UK becoming the first country in the world where electronic health records can be combined with knowledge of patients' genetic make-up, which would be sequenced at birth in the future (Telegraph 2013). In time, and with the expansion of the 100 k Genome Project, whole-sequenced genomes will be connected to care.data through HSCIC (Nuffield 2015, p.108).

The last major development concerns the Department of Health disclosure that personalised medicine will be incorporated into mainstream NHS healthcare, from 2017. As personalised medicine requires whole-sequenced genomic data, and will soon be integrated into mainstream healthcare – this is another driver in the normalisation of whole genome sequencing. Here, we can see how the Governments ambitions to become a leader in personalised medicine and genomics, intersect.

2.1 Legislative Changes

In addition to these infrastructural developments, legislative changes and changes to the NHS constitution were to be made. The first installment was Section 251 of the NHS Social Care Act (2006). Granted by the Secretary of State for Health, it circumvents the common law of confidentiality, Article 8 of the Human Rights Act (2008) and the Data Protection Act (1998). The Health & Social Care Act (2012) goes a step further. It provides access to patient-identifiable data available, via the NHS Commissioning Board and the HSCIC, without the need to seek approval from the Secretary of State for Health and, in the case of medical research, by a research ethics committee or the Health Research Authority. Accessing identifiable patient data for purposes beyond direct care required making changes to the NHS constitution, which, David Cameron PM affirmed, would make every NHS patient a "research patient", and their medical details, "opened up" to private healthcare firms (BBC 2011).

Slowly but surely, the landscape of information governance was changing. In 2013, and motivated by the conviction that the risk-averse attitude towards sensitive patient data had ultimately been at the expense of public benefit (Information: To share or not to share? The Information Governance Review, The Department of Health (DoH) 2013, p.62). Dame Fiona Caldicott was invited by the Secretary

of State for Health to lead an independent panel of experts to conduct a major independent review of information governance. Caldicott2, was to ensure that there is "an appropriate balance between the protection of patient information and the use and sharing of information to improve patient care" (IBID p.6). The review concluded that data collected during the course of care and treatment should be used to support research (IBID p.13). Further, personal confidential data or data that has been de-identified, but still carries the risk of re-identification, should only be accessed for any purpose other than direct care – under certain conditions, namely, in accredited environments, 'safe havens' (IBID p.67). A consultation on setting up a number of safe havens across England was launched in the summer of 2014.

3 The Changing Landscape of Information Governance

The first information governance review (1997) set out conservative principles to ensure the maintenance of patient confidentiality (Nuffield 2015, p.37). Caldicott2 marks a shift in focus from autonomy, privacy and confidentiality, to a significantly wider range of values. On Nuffield's (2015 p.23) analysis, the value of these data initiatives is three-fold. These are, cost reduction and resource management through the better use of business intelligence. Further, since greater access to data is likely to improve our understanding of the use of treatments and medicines, providing access to data held across the NHS has the potential to result in improvements to the healthcare system. Lastly, these initiatives will also help to improve the practice of medicine, and bring about advances in science, which is likely to benefit NHS users. This, then, presents a case for the inclusion of these considerations into the governance of sensitive, personal medical and genomic information. In other words, there are other values than privacy at play and the governance of personal medical and genomic information should change accordingly.

A further consideration is that, the changes that have been documented have afforded commercialisation incomparably greater force in the governance of medical and biomedical data. This development is tied to a more general trend in which the economic paradigm has come to dominate policy relevant to the biosciences in the UK (Nuffield 2012, p.xxiv). The drivers behind this trend were anchored in a need to respond to the global financial crisis, beginning in 2007, and the subsequent economic downturn, which led governments to foster economic growth by focusing on existing assets. In the UK, this focus has fallen on, among other things, the exploitation of public sector data, IT innovation, and the strong research base in the biosciences (Nuffield 2015 p.22). Changing economic conditions certainly played a part in motivating the introduction of these data initiatives. However, private sector interest in unlocking the considerable dataset that is held across the NHS and creating the capacity to store whole-sequenced genomes, and to link these to clinical records and relevant phenotypic data has been a key factor. This is evidenced by the extracts of submissions to the Select Committee on science and technology (2000, 2007) that were considered at the start of this section.

Pinpointing the causative factors in the relaxation of the governance of personal medical and genomic information is no easy task. Ultimately, the reasons behind this are multi-layered. Economic factors and a vested interest in securing a competitive advantage have been key drivers, but changes to values in research ethics had already taken place. The promise of genomics has been the determining factor in the shift from autonomy, privacy and confidentiality (Knoppers and Chadwick 2005; Lunshof et al. 2008). Whether the data initiatives that have been considered are a determination of economic policy or not, the relaxation of the regulatory regime that has taken place is, nevertheless, something that genomics unavoidably requires. In other words, if it hadn't happened now then it would have happened further down the line. Though this may be true of genomics, the same cannot be said of personal medical data held across the NHS.

4 The 'Value Impact' of Genomics

Two decades ago, autonomy, privacy, justice and equity were the norms that framed human genetic research internationally (Knoppers and Chadwick 1994). During the last decade or so, the understanding of the complexity of genetic factors in common diseases and of the familial and socio-economic impact of genetic information and genetic tests, together with the concomitant expansion of public participation in policy-making, have given rise to new trends in ethics (Knoppers and Chadwick 2005, p.75). 'Reciprocity', 'mutuality', 'solidarity', 'citizenry' and 'universality' have come to prominence (Knoppers and Chadwick 2005, p.75). Though genomics has had a real impact on values (Knoppers and Chadwick 2005), this ought not to be considered a problem (Lunshof et al. 2008). There are times, when it is necessary to revise even the most basic ethical models (Chadwick and Berg, 2001, p. 318). Indeed, the realities of genomics instantiate such a requirement (Lunshof et al. 2008). When it comes to genomics, upholding medical confidentiality, the principle of 'informed consent' and autonomy would ultimately be prohibitive to genomics. A more congenial model of consent for genomics is 'open', 'broad' or 'generic' consent (Lunshof et al. 2008).

With an open model of consent, by consenting to their whole-sequenced genome, and extensive phenotypic information being used, any control over subsequent uses of their genomic data is relinquished. Furthermore, since it is technically impossible to securely anonymise a person's genome, this implies the foregoing of confidentiality that forms part of the principle of 'informed consent' (Lunshof et al. 2008, p.252). In this light, the 'value-impact' (Lunshof et al. 2008) that can be traced to genomics falls on a network of values – at the outer layer is the principle of 'informed consent.' 'Informed consent' has as its corollaries, privacy or confidentiality. And at a deeper level, these values are connected to informational and decisional autonomy. And more foundationally still, we have the principle of 'respect for persons' or autonomy. Though autonomy remains prominent, it now has to compete with 'reciprocity', 'mutuality', 'solidarity', 'citizenry' and 'universality'

(Knoppers and Chadwick 2005, p.75). Any loss in autonomy, however, is redressed by the expansion of public participation in policy making (Knoppers and Chadwick 2005, p.75). In other words, our wishes and preferences are still being considered, but in the development policy-making, rather than the autonomy that is conferred to participants as is required by 'informed consent.'

5 Changing Perspectives

As well as charting a shift in values, this section evidences a shift in opinion. Initially, and certainly in academia, there was a calm acceptance that genomics would mean a departure from the principle of 'informed consent'. Recently, though, the case for a return to the primacy of 'respect for persons', which underpins 'informed consent', has been made (Nuffield 2015). In contradistinction to mainstream bioethics, which takes normative questions as its starting-point, ignores broader, relational questions, Nuffield (2012) focuses on the social, economic and the political environment in which new technologies are emerging and policy and regulatory decisions are being taken. Though communitarian lines of thought may have become more influential in bioethical literature in the last decade (Chadwick 2011), recent policy developments in genomics have tended to narrow down decisions, focusing on such things as delivering economic growth, rather than taking into account broader concerns, derived from a multiplicity of public perspectives, about the value of social life and the public good (Nuffield 2015, p.xix). Contra the view that policy developments in genomics have been accompanied by public participation in policy-decisions (Knoppers and Chadwick 2005, p.75), sectional interests, rather than a plurality of perspectives, are likely to be the determining factor in policy-developments (Nuffield 2012). The lack of transparency, openness and inclusivity that has marked the HSCIC and the 100 K GP provides empirical support for the perspective set forth by Nuffield (2012).

Nuffield (2012) articulate the makings of a new paradigm that is specifically tailored to meet the novel and complex challenges of the biosciences. The biosciences, and advanced biotechnology in particular, have the potential to transform or displace existing social relations and practices or indeed, give rise to new capabilities that did not previously exist or are yet, unimagined (Nuffield 2012, p.40). Due to epistemic uncertainty about the range of possible outcomes that may occur, the biosciences could give rise to significant benefits but also – due to misuse, unintended consequences and uncertainties – be profoundly harmful (Nuffield 2012, p.40). Lastly, as there is a lack of agreement about the wider consequences, implications, meanings and range of possible outcomes, the biosciences are inherently ambiguous (Nuffield 2012, p.40). Cognisant of the uncertainty, ambiguity and transformative potential of the biosciences, a 'broad view' must be taken when making policy and regulatory decisions (Nuffield 2012, 2015). What they propose marks a departure from biotechnology policy which has hitherto brought together two opposite elements. State intervention, in which 'experts' decide where to concentrate the resources

available for technology development, and market mechanisms, which aggregate people's preferences about whether new technologies are desirable or undesirable (Nuffield 2012).

As a means to aggregate social preferences, we cannot rely on markets to guarantee that resources will be allocated in ways which are instrumental to important social concerns (Nuffield 2012, p.1; Nuffield 2012, p.178). Further, modern societies are arranged in such a way that, policy decisions are likely to be heavily influenced by sectional interests rather than those of society as a whole (Nuffield 2012, p.66). Introducing a social value, through public engagement, as a third element in the shaping and selection of policy decisions in the biosciences can foster a more socially responsible approach to policy and governance, and cultivate a mode of thinking about policy and governance as a matter of social choice (Nuffield 2012, p.xxvi). The justification for taking a 'broad view' of bioscience policy and regulation would appear to be three-fold. Affording the public a central role in policy and governance decisions might offset the influence of market forces and sectional interests and address any democratic deficit. Secondly, any harm that is likely to occur from their introduction into society is likely to affect the majority of people – and society, at many different levels, and in non-trivial ways (Nuffield 2012, p.xx). As such, there is a significant public interest in the biosciences.

The second consideration would appear to be encompassed by the 'democratic imperative'. That is to say, in a democratic society, the involvement of citizens in policy-making in areas likely to have a demonstrable impact on everyday lives is something that ought to be striven for (Sturgis 2014, p.41). The third concerns the supposition that including a more diverse range of perspectives and values can widen the interaction and scope for reflexivity in the development of policy (Wynne 2011) and thereby, act as a counterbalance to technical or economic interests (Nuffield 2012, p.xxi). Thus, giving rise to potentially more robust policy decisions (Nuffield 2012, p.xxi; xix), that are oriented towards more socially beneficial ends (Stirling 2008, p.39).

6 The Locking-in of the Economic Paradigm in the Biosciences

The following section focuses on novel developments within the governance of medical and biomedical 'Big data'. Though advances in information technology, bioinformatics and genomics have led to a change in attitude towards the secondary uses of data (Nuffield 2015), the commercialisation of pseudonymised or identifiable data raises ethical considerations, which have not been subject to adequate levels of public consultation or debate. There is an important difference between the extraction of sensitive personal medical and genomic data for purposes related to improving healthcare, science and medicine, and the use of these data to further the knowledge economy. However, policy in the biosciences and health sector have had the aim of promoting and, to lock in, the exploitation of data as a way

of boosting the economy in the short term, and to establish the conditions for improved and more cost-effective treatments and services in the long term (Nuffield 2015, p.43). Short-term economic motives, however, are likely to have long-term consequences. Though these developments are taking place with insufficient regard for their social and ethical implications, we can be sure that the policies and regulatory frameworks currently being enacted will shape future decisions. It is therefore likely that the move towards commercialisation results in the locking in of commercialisation in the biosciences.

The language used in relation to medical and genomic data by policy-makers evidences a shift in the terms that are appropriate to the domain of personal medical and biomedical data, thus rendering the locking in of commercialisation in the biosciences more likely. In regard to "research opportunities and mainstream use of genomic medicine across the NHS", NHS England affirmed that these could make a "major contribution to wealth creation and economic growth in this country" (NHS England 2014). And the body tasked with providing ethical guidelines in the 100,000 Genome Project, the Ethics Advisory Group, refer to the projects "potential to bring real benefits to individual patients and their families, to the NHS more broadly, and to the UK economy" (Genomics England 2014). This tendency towards commercialisation is linked to two trends within genomics and one, more all-encompassing trend outside of it. Indeed, the more general trend is left out of the analysis of the economic factors that were among the driving forces behind an economic policy which focuses on the biosciences, given by Nuffield (2012). At a micro-level, there has been increasing prevalence of investments into biomedical research from private enterprises. Consequently, the principles and language of the business-world have become more influential within genetic research (Chadwick and Hedgecoe 2008). The growing influence of the private sector has led to what Caulfield and Williams-Jones (1999) described as a 'pro-commercialisation environment' in genomics and biobanking.

At a macro-level, the biosciences are not immune to wider socio-economic tendencies, such as those of liberalisation, privatization and deregulation. What is particularly distinctive about this wider trend, often captured by the term neoliberalism, is that it 'desacralizes' institutions and public services, such as education and healthcare, which were previously protected from market forces and influence (Mudge 2008, p.704). Though the commercialisation of the biosciences may indeed be linked to wider tendencies, the commercialisation of human genomic material remains controversial. For instance, the Universal Declaration on the Human Genome and Human Rights (UDHGHR) (1998) urges that the human genome in "its natural state shall not give rise to financial gains." In the UK the commercialisation of genomics now forms part of economic policy. Once the biosciences are framed in terms of commercial potential, it then becomes normal to think about this sphere in commercial terms, making further commercialisation easier. As the forces of commercialisation take hold, the purposes for which medical and genomic data are used by commercial entities and Government, for example, are likely to widen.

7 Reflexivity

Consideration will now be given to the conditions in which policy decisions in the 'new genetics' are taken. The analytical construct through which these policy decisions are explicated is that of a 'frame', which derives from the work of Goffman (1986). There are many social processes that operate in the real world that have the effect of closing down the plurality of frames that may be applied in the governance of the biosciences, and in ways which that serve sectional interests (Nuffield 2012, p.51). Since bioscience policy in the UK has become increasingly framed in terms of a paradigm of economic growth, other values, which are perhaps harder to quantify, are being obscured (Nuffield 2012, p.xxiv). An open and reflective approach, derived from a plurality of perspectives, however, can open up a range of alternatives that, all too often, are absent in current technology governance, thereby illuminating contingencies, such as obscured assumptions or constraints (Nuffield 2012, p.xviii).

Once established, and with momentum, technologies can become locked-in, thereby foreclosing alternative pathways (Nuffield 2012, p.16). Public engagement at an early stage, that moves 'upstream' (Wilsdon and Willis 2004), can prevent future trajectories of technology being prematurely closed down (Stirling 2008). Though, it is the public interest that a balance is struck between the pro-R&D agenda that is driving a permissive regulatory regime in the UK, and sensitivity towards social and ethical concerns that public engagement can potentially alert us to, the public relations strategy employed by successive governments' has meant that the conditions for public deliberation about the significance, risks and wider implications of these initiatives have not been in place. Indeed, the Government has made the locking in of commercialisation in the biosciences more likely because the conditions for meaningful discussion were thwarted by the deliberate obfuscation of the associated risks and wider implications. Though it is only by deliberating over such things as the appropriate model of consent and its attendant risks, that society can more clearly debate the benefits and risks that are entailed by each course of action (Caulfield et al. 2003, p.4), an articulation of the risks requires access to all relevant information. In this regard, the Government has deliberately opted to distort public understanding of the level of access and the wider trajectory of policy developments in the UK, thereby inhibiting intelligent debate.

8 Democratic Deficit and the Conditions for Meaningful Public Discourse

Let us cast our mind back to the submission to the Select Committee on science and technology by Smith Kline, in 2000, and the perceived need for "the political will to tackle the issue of public acceptability"; and to "address public concern over many of the ethical issues... in the context of medical privacy, use of anonymous data

and consent issues." The strategic use of the term anonymisation by Government and stakeholders has meant that vital discussion of the risks, and the challenge of securing public acceptability, has been circumscribed. In 2007, the Association of the British Pharmaceutical Industry recommended that "future systems should support use of patient level data via an opt-out patient consent protocol (2007)" Soon after, default opt-in became the *de facto* mechanism for 'Summary Care Records', the electronic medical record system that was part of 'Connecting for Health', and through the 'Secondary Uses Service', every NHS user became a research subject. Throughout, as Government and other stakeholders maintained that only anonymised personal data would be shared, this had the effect of impoverishing the debate and made taking a 'broad view' on these developments seem unnecessary.

The strategy employed with the Secondary Uses Service was also present when introducing the new wave of data initiatives; the CRPD, the HSCIC and the 100 k GP. David Cameron PM insisted that these initiatives will not "threaten privacy", and only "anonymous data" will be made available to make new medical breakthroughs (BBC 2011). The then Health Secretary, Andrew Lansley, claimed that individual records would not be shared in a way in which individuals could be identified. Dame Sally Davies, Chief Medical Officer and Chief Scientist at the Department of Health, insisted that as safeguarding confidentiality of patients is a priority, all data is anonymised and patients can opt-out (Sample 2012). It came to light, however, via *The Guardian*, that private health firms including Bupa had been given approval to access 'sensitive' and 'identifiable' patient data held on the HSCIC (Ramesh 2014). Further, contradicting prior public announcements, and what Genomics England claim on their website, a Freedom of Information disclosure by the Department of Health regarding the level of access to whole sequenced genomic and clinical data of participants in the 100 k GP confirmed that data made available to third parties, including commercial entities, would not be anonymised, but rather, "pseudonymised". This has profound implications.

Anonymised data is stripped of anything that would permit the identification of the individual in question from the data. By contrast, pseudonymised information – in the DoH's own words, contains "age or age range" and "wider geographical information". The information made available to third parties will also include clinical data pertaining to an individual's medical history, potentially spanning decades. Combining this with the wealth of 'big data' held on databases and available online, it may then be possible to identify those participating in the 100 k GP. In light of the particular strategy taken up by successive governments when implementing these data initiatives, the cautiousness and moderate pessimism of Nuffield (2012) appears justified. There have been many occasions in which the public might have been given a more central role; Caldicott2 is a case in point. Through democratic consultation with all relevant stakeholders, the information governance review might have provided a timely, non-partisan assessment of the information governance regime. However, prior to the publication of Caldicott2's recommendations in March 2013, the government's 'Strategy for UK life sciences 2012' had already made it clear that there would be a shift to 'a more progressive

regulatory environment.' Caldicott2 was therefore never going to be able to meet its aspirations to give the public a stake in deciding whether or not information would be shared.

The lack of transparency and inclusivity is in stark contrast to UNESCO's International Declaration on Human Genetic Data IDHGD (2003), which calls upon nation-states to "endeavour to involve society at large in the decision-making process concerning broad policies for the collection, processing, use and storage of human genetic data."[1] A further contributing factor in what had become a narrowly focused debate on the 100 k GP, is the view that "genetic information should not be treated any differently from other forms of information, and genetic information in itself is not always identifiable" (Information: To share or not to share? The Information Governance Review (Caldicott2) 2013, p.59). Relevant legal and expert, international, regional and domestic normative guidance and legal instruments, however, acknowledge the sensitivity of genetic material and do not equate genetic material or data with more trivial types of information (Knoppers 2005, p.10–11). For instance, UNESCO's International Declaration on Human Genetic Data (2003) (IDHGD), states that since genetic data can be predictive of genetic predispositions, and thus has the capacity to impact on relatives, such as offspring, extending over generations, it has a "special status" (2003, p.41).

Depending on the context in which it is used and how it is linked to other related information, biomedical data can be extremely 'sensitive' (Nuffield 2015, p.19). The sensitivity of human genetic data requires that appropriate levels of protection are in place (UNESCO 2003, p.41). More specifically, when research with human genetic data is undertaken, 'informed consent', that is, the prior, free and informed consent of the person whose data is being used, must be obtained (UNESCO 2003, p.43). The Universal Declaration on Bioethics and Human Rights UNESCO UDBHR (2005) is the first of its kind in that member states agreed to follow the provisions set out by the declaration; and Article 6 makes the need for consent in any scientific research an explicit requirement.

8.1 Societal Tendencies and the Normative Grounds of a Deliberative Approach

Relevant international normative guidance and legal instruments maintain that human genomic data has a special status, and policy decisions concerning its usage must be democratically accountable. The initiatives that have been considered have not treated genetic data with particular sensitivity or adhered to the normative recommendations that policy ought to proceed along democratic lines. Indeed, the way policy developments have unfolded is very instructive. They exemplify how

[1]Except for an "important public interest reason in cases restrictively provided for by domestic law consistent with the international law of human rights or where the prior, free, informed consent of the person concerned."

norms crystallise in modern societies – evidencing, in particular, a democratic deficit in developments of great import to society and to people's interests. A corollary of this democratic deficit is that public deliberation about the significance, risks and implications of novel innovations within science and technology is often circumscribed. This brings us to the crux of the problem; if this is symptomatic of a general societal tendency, can we expect normative prescriptions for democratic accountability to be efficacious? On one hand, we have the normative requirement regarding democratic accountability in bioscience policy and regulation. On the other, there are the realities of liberal, democratic, free-market, societies upon which such normative stipulations are supposed to bear. Consideration will firstly be given to the tensions found therein. Following this, an objection to bioscience policy and regulation being shaped by market thinking and sectional interests, will be grounded in a normative theory of democratic deliberation and an explication of the capacity modern societies have for democratic deliberation.

An instructive way illustrating these tensions is to adopt a macro-level perspective of society, which focuses, in particular, on societal tendencies. If democratic deliberation is to be taken seriously, then it is necessary to acknowledge two things; the expansion of the reach of market orientated thinking, and the emptiness of public discourse (Sandel 2012, p.11/12). In moving from having a market economy to being a market society, market orientated thinking is now not only confined to material goods, but extends into aspects of life traditionally governed by non-market norms (Sandel 2012, p.10). Further, questions of great importance tend to be framed in purely economic terms (Sandel 2012, p.10). Like Michael Sandel, Jürgen Habermas (2015, p.p.71/81) is of the view that modern societies are characterised by a tendency towards de-democratisation which has led to the balance between politics and the market coming out of sync. This evolution into a market society necessitates public discourse about the limits of the market and the types of goods and practices that shouldn't be subject to its logic. Indeed, questioning and evaluating the proper reach of the market on a case by case basis, and through public dialogue, has the potential to re-invigorate our politics (Sandel 2012, p.6).

If the policy developments of concern to this chapter can be shown to be ethically problematic, then the following questions will have to be answered. Is there moral warrant to afford priority to the public interest rather than market mechanisms and sectional interests? And can a dialogically reached agreement in regard to innovation in science and technology, in general, and the biosciences, in particular – which reflects the wider interests of society – be reached? The communicative competence of members of modern societies provides the starting-point from which we can appreciate the potential inherent to modern societies to reach such an agreement. What distinguishes modern societies is that the structures of communication are rationalised (Habermas 1984). Under conditions of rationalisation, communicative utterances correspond to three different worlds, and each world has its own distinct criteria for validity. In the objective world it is propositional truth, in the social it is normative rightness, and the subjective is subjective truthfulness (Habermas 1984, p.75). The thesis that the aim of speech

is not only validity, but also, acceptability (Habermas 1987, p.p.91/107) leads us to a crucial step in Habermas' theory: communicative rationality.

As communicatively achieved understanding must be based on reasons that are rationally acceptable to others – rational, in so far as we might have to defend the claims we make and convince others that our utterances are sincere, true or right – communicatively achieved undertsanding is rational (Habermas 1984, p.p.13/51). The hallmark of a validity claim is that an appeal to reasons or grounds that are acceptable to others is made, which is, in principle, always open to criticism and may require further justification. When validity claims are subject to rational scrutiny, a shift takes place from everyday communication to a reflective form of communication: discourse. The realm of discourse furnishes the conditions for communicative rationality, and it is these conditions that provide the basis for a communicatively rational agreement between individuals, groups or members of a society. Discourse is an implicitly rule-governed activity, requiring the inclusion, freedom and equality of all those that participate (Ingram 2012, p.83). Among its implicit norms is the stipulation that nobody who could make a relevant contribution may be excluded. Further, all participants are afforded equal opportunities for participation. Participants must be sincere, and communication must be free from internal and external compulsion, so that validity claims should only be accepted by the unforced force of better reasons (Finlayson 2000, p.13). In addition to these, Habermas includes three further rules. Everyone is allowed to question any assertion; and everyone is allowed to introduce any assertion into the discourse. Further, everyone is allowed to express his attitudes, desires and needs (Finlayson 2000, p.13).

In the proceeding pages, an account of the transition from everyday communication to the communicative rationality inherent in discourse has been given. What happens when we enter the realm of discourse has definite implications for ethics. In the absence of religious or metaphysical worldviews that claim to be immune from criticism or scrutiny, the practical orientations of rationalised societies are grounded in the reflexive forms of communication (Habermas 1996, p.98). In other words, the reflexive form of communication – discourse, provide a source of normativity for modernity. It is through an adherence to the norms of discourse and the attainment of a rationally motivated consensus that any resulting agreement, decision or norm can be considered valid (Finlayson 2000, p.9). On this account, morality is conceived as an open and inclusive process of argumentation, in which everyone is afforded equal participation and are free to introduce and question claims. The underlying structures of communication generate two moral principles.

The 'discourse principle' stipulates that that "only those norms can claim to be valid that meet with the approval of all affected in their capacity as participants in a practical discourse" (Habermas 1990, p.93). The 'universalization principle' states that "all affected can freely accept the consequences and the side effects that the general observance of a controversial norm can be expected to have for the satisfaction of the interests of each individual" (Habermas 1990, p.93). For any decision or agreement to be considered morally binding, such as whether to proceed with genome sequencing on a national scale, then all those that stand to be affected

by it must have considered and accepted the consequences, and have agree to it. Since communicative rationality admits of the potential to achieve, sustain and renew a consensus, governance decisions, such as: who should have access to the 100 k GP data, and for what purposes, have the potential to be the product of a communicatively rational consensus.

The point of this exposition of the structures that make rational orientations of action possible for individuals, groups and society's (Habermas 1984, p.44), is to demonstrate that key decisions in the biosciences can be both reflexive, and reflect general rather than sectional interests. The challenge, however, given the way policy developments of concern to this chapter have unfolded, is actually having a practical discourse in which a plurality of perspectives and value-commitments are considered. A macro-level analysis has great explanatory power regarding the societal tendencies that are driving a narrow focus on an economic paradigm, short-termism and the privileging of sectional interests over those of democratic deliberation and a plurality of perspectives. A communicatively rational approach has the potential to foster adequate discussion of the wider significance of bioscience policy, and the associated risks and implications of governance decisions. Before this line of thought is considered further, consideration will be given to the risks associated with genomics.

8.2 A Clarification of the Risks

The supposition that genetic information is analogous to other, more trivial types of information (Department of Health 2013) is connected to the view that its introduction into society doesn't warrant special consideration. This view is reinforced by the discussion of the risks in academic literature, which centres upon the potential for re-identification and the harms that may ensue, and leaves out deeper analysis of the social, legal and ethical dimensions of the relaxation of the governance of medical and biomedical 'Big data'. Focusing on a narrow conception of risk belies the complex social dimensions of the biosciences. Since the relaxation of the standard of consent in genomics opens up the possibility that the information on whole-sequenced genomes could be used for a variety of purposes, a departure from 'informed consent' is less trivial than what one might suppose. Then, there is the embedding of the value of commercialisation into the information governance regime. The implications of both of these developments are the focus of this section.

One of the central observations of this chapter is that we are witnessing a shift from a rights-based approach to the adjudication of competing claims, in which benefits to the economy, for example, are seen as goods to be balanced with a data subject's right to privacy and confidentiality. As result of the changes that have been documented, The Human Rights Act (1998), The Data Protection Act (1998) and the Common Law Duty of Confidentiality have to compete with other values in the governance of sensitive medical or biomedical data. The relaxation of the governance of sensitive information affords new levels of access and opportunities

for commercialisation for the private sector. It engenders new responsibilities for the Government, too. If indeed democratic ideals are valued, then moving beyond informed consent must be accompanied by public debate (Mittelstadt and Floridi 2016, p.13), in particular in regard to uses of genomic data, by whom and for what. Before this line of thought is developed further, it is necessary to consider the academic discussion on risk.

Lunshof et al. (2008) explore the threats to privacy and confidentiality as a result of the re-identification after de-identification, through such things as gaining access to publically available data, inferring the phenotype from the genotype by identifying information in DNA such as height, hair or skin colour; identification through the DNA of a first-degree relative; obtaining any genomic data that is in the public domain with a name which can be used to identify any anonymized genomic data set; identifying a phenotype through imaging techniques for reconstruction of facial features; hacking into computer systems, and lastly, theft or loss of data-storage devices or a laptop (Lunshof et al. 2008, p.406). The analysis of risks given by Heeney et al. (2011) widens the parameters somewhat. They employ the term 'data intruder' for anyone that is motivated to know more about the attributes or identity of a data subject by acquiring available information, such as pseudonymised data. The 'data environment' of genomics includes many people that have these motivations for reasons which include using genetic information in marketing, insurance or employment decisions (Heeney et al. 2011). The common thread in the literature is that risks are explicated in terms of identification after de-identification, by illicit or non-authorised means. Little is said, however, about the range of potentially legitimate uses or those that may be within the bounds of law.

The HSCIC's Public Assurance Director acknowledged that there was great ambiguity regarding who might be able to access patient data, "a government department, university researcher, pharmaceutical company or insurance company" (Guardian 2014) could all potentially be granted access. A further concern is that, in the UK, legislative and normative guidance regarding acceptable and non-acceptable uses of whole-sequenced genomes is indeterminate. In regard to the model of consent employed in the 100 k GP, the Department of Health has claimed that such is its ambition to stay at the "forefront of genomic research, it is impossible to inform patients at the outset of the potential ways in which their genome might be used." The body tasked with providing ethical guidance regarding potential uses of whole-sequenced genomes maintains that it would be impractical to place restrictions on the research undertaken on genomic data, such as limiting it to 'non-commercial research' (Ethics Advisory Group 2013).

The infrastructure is currently being laid out with private sector involvement for what looks to be an expansion of the 100 k GP, potentially on a national scale: a '50 million Genome Project'. A disclosure by the Department of Health (DoH) to Ethics and Genetics confirmed that, a "decision will be made by the Secretary of State for Health following discussions with a range of interested parties." A concern, however, is that equal weight is unlikely to be given to a range of interests, and that sectional interests – which have been a determining factor from the beginning – will continue to be prioritised. Genomics England has

announced that it will be collaborating with GlaxoSmithKline amongst others, and with the introduction of personalised medicine into the NHS from 2017, the level of involvement of commercial entities is set to increase. In 2013, Patrick Chung of 23andMe claimed that they "will make money by partnering with countries that rely on a singlepayer health system" (Fast Company 2013). By genotyping "everyone in Canada or the United Kingdom… the government is able to identify those segments of the population that are most at risk for heart disease or breast cancer. You can target them with preventative messages… 23andMe has been in discussion with a bunch of such societies" (Fast Company 2013). A Freedom of Information disclosure revealed that, during meetings with the Department of Health, Google and 23andMe expressed an interest in the 100 k GP data.

9 New Powers and Novel Challenges

Should it be developed, the significance of a national Genome Project reaches far beyond improving health, the economy, and advancing science. In 2013, and owing to the bravery of Edward Snowden, the details of a mass electronic surveillance programme by the US National Security Agency came to light. As whole-sequenced genomes are a wealth of information, sequencing genomic information on a national scale would appear to increase the capacity for surveillance and control. The Snowden revelations, along with single-issue pressure groups such as Medconfidential, have played an important part in the postponement of, and uncertainty surrounding, the implementation of Care.data. By contrast, the 100,000 GP hasn't faced either scrutiny or a level of analysis commensurate with its significance. Economic interests and embedded power structures account for the narrowing down of the way bioscience policy, in general, and the 100 k GP, in particular, has been framed. This has meant that subtler, longer-term concerns about the social impact of these developments are being overlooked.

The public is capable of showing sensitivity towards the moral nuances of the biosciences. The forces of commercialisation are blind to the morally relevant differences between, for instance, a commercial model of germ-line gene editing and therapeutic uses of gene editing to eliminate disease causing alleles, for example. Fully transparent and inclusive public consultation in regard to regulatory decisions and the trajectory of policy in the biosciences, could keep in check potentially malign appropriations of genomics by Government; and, further, ensure that private sector interest in commercialising genomic data remains conducive to maximising the potential of this new frontier in science and medicine, without shaping social conceptions of the biosciences in ways which could have substantive implications for society. Lastly, when decisions of considerable importance are to be taken, such as the decision to proceed with germ-line gene editing, public referendum is required.

10 Conclusion

Whilst Nuffield (2012) articulate the makings of a new paradigm that has been specifically tailored to meet the novel and complex challenges of the biosciences, wider societal tendencies are not taken into account. Developments in the UK demonstrate that the macro-level perspective argued for herein is instructive, but is also evidence of the challenges democratising policy entails. Neoliberalism is eroding the substantive citizenship (Brown 2006, p.690) that a public-centred approach to bioscience decision-making depends upon. Nuffield in fact recognise that it may not sufficient to reconfigure decision-making procedures, rather, it might ultimately be necessary to alter behaviours (2012, p.2). However, this vital component is not developed further, and appears to be evolutionary in nature. Assuming that the imbalance between the market and democracy is redressed and public participation in bioscience policy becomes more substantive, many important decisions will already have been taken. The possibility that policy and regulatory decisions will become locked in, and the foreclosing of alternative pathways, calls for a sense of urgency. If the right conditions are in place, however, a democratic and deliberative approach could steer policy and regulatory decisions towards the interests of society as a whole. Prevent malevolent use of the biosciences by Government or the private sector, and allow the general public and other stakeholders to determine how we think about this new frontier in science, and the challenges it brings to bear.

References

Brown, Wendy. 2006. American Nightmare Neoliberalism, Neoconservatism, and De-Democratization. *Polit Theo* 34(6): 690–714.

Caulfield, Timothy. 1999. Regulating the commercialisation of human genetics. Can we address the big concerns? In *Genetic information. Acquisition, access, and control*, ed. A. Thompson and R. Chadwick, 149–162. New York: Kluwer Academic/Plenum Publishers.

Caulfield, Timothy, Ross Upshur, and Abdallah Daar. 2003. DNA databanks and consent: A suggested policy option involving an authorization model. *BMC Medical Ethics* 4: 1.

Chadwick, Ruth. 2011. The communitarian turn: Myth or reality? *Cambridge Quarterly of Healthcare Ethics* 20(4): 546–553.

Chadwick, Ruth, and Kåre Berg. 2001. Solidarity and equity: New ethical frameworks for genetic databases. Nature review. *Genetics.* 2: 320.

Chadwick, Ruth, and Adam Hedgecoe. 2002. Commercialisation of the human genome. In *A companion to genethics*, ed. J. Burley and J. Harris. Oxford: Blackwell.

Department of Health. 2012. Building on our inheritance: Genomic technology in healthcare. https://www.gov.uk/government/publications/genomic-technology-in-healthcare-building-on-our-inheritance. Accessed 11 June 2014.

Department of Health. UK Gov. 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/210830/ethics_advice_letter_to_CMO.pdf. Accessed 24 Nov 2014.

Ethics Advisory Group. 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/210830/ethics_advice_letter_to_CMO.pdf. Accessed 7 Mar 2015.

Everyone 'to be research patient', says David Cameron. 2011. *BBC News*, December 5.

Evidence submitted by the Association of the British Pharmaceutical. 2007. Industry Select Committee on Health Written Evidence. House of Commons Select Committees. http://www.publications.parliament.uk/pa/cm200607/cmselect/cmhealth/422/422we05.htm. Accessed 27 June 2014.

Finlayson, Gordon. 2000. Modernity and morality in Habermas's discourse ethics, inquiry: An interdisciplinary. *Journal of Philosophy* 43(3): 319–340.

Genomics England. 2013. http://www.genomicsengland.co.uk/genomics-england-launch/. Accessed 15 Mar 2015.

Genomics England. 2014. http://www.genomicsengland.co.uk/prof-mike-parker-says-no-ethical-issues-are-off-the-table-as-gel-considers-its-approach-to-patient-consent/. Accessed 10 Mar 2014.

Goffman, Erving. 1986. *Frame analysis. Reprint*. Boston: Northeastern University Press.

Gray, Richard. 2013. Children could have DNA tested at birth 2013. *Telegraph*, December 8.

Gunson, D. 2010. The philosophical foundations of the discourse society. In *Oppositional discourses*, ed. M. Huspek. Oxford: Routledge.

Habermas, Jürgen. 1984. *The theory of communicative action, vol. 1: Reason and the rationalization of society*. London: Heinemann.

Habermas, Jürgen. 1987. *The theory of communicative action, vol. 2: Lifeworld and system: A critique of functionalist reason*. Oxford: Polity Press.

Habermas, Jürgen. 1990. *Moral consciousness and communicative action*. Cambridge: Polity Press.

Habermas, Jürgen. 1996. *Between facts and norms*. Cambridge: Polity Press.

Habermas, Jürgen. 2015. *The lure of technocracy*. Cambridge: Polity Press.

Heeney, C., N. Hawkins, J. de Vries, P. Boddington, and J. Kaye. 2011. Assessing the privacy risks of data sharing in genomics. *Public Health Genomics* 14(1): 17–25. doi:10.1159/000294150.

Ingram, David. 2012. *Habermas: Introduction and analysis*. Ithaca: Cornell University Press.

Knoppers, Bartha. 2005. Biobanking: International norms. *The Journal of Law Medicine and Ethics* 33(1): 7–14.

Knoppers, Bartha, and Ruth Chadwick. 1994. The Human Genome Project: Under an international ethical microscope. *Science* 2035–2036.

Lunshof, Jeantine E., Chadwick Ruth, B. Vorhaus Daniel, and M. Church George. 2008. From genetic privacy to open consent. *Nat Rev Genet* 9(5): 406–411.

Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. The ethics of Big Data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.

Mudge, Stephanie. 2008. What is neo-liberalism? *Socio-Economic Review* 6(4): 703–731.

Murphy, Elisabeth. 2013. Inside 23andme founder Anne Wojcicki's $99 DNA revolution. Fast Company. http://www.fastcompany.com/3018598/for-99-this-ceo-can-tell-you-what-might-kill-you-inside-23andme-founder-anne-wojcickis-dna-r. Accessed 13 May 2014.

NHS England. 2014. http://www.england.nhs.uk/2014/07/03/genomic-medicine/. Accessed 19 June 2015.

Nuffield Council on Bioethics. 2012. Emerging biotechnologies: Technology, choice and the public good. http://nuffieldbioethics.org/wpcontent/uploads/2014/07/Emerging_biotechnologies_full_report_web_0.pdf. Accessed 21 Mar 2015.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*. http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf. Accessed 21 Mar 2015.

Ramesh, Randeep. 2014. *NHS patient data to be made available for sale to drug and insurance firms. Guardian*, January 19.

Sample, Ian. 2012. *NHS patient records to revolutionise medical research in Britain. Guardian*, August 28.

Sandel, Michael. 2012. *What money can't buy: The moral limits of market*. New York: Farrar, Straus and Giroux.

Sturgis, Patrick. 2014. On the limits of public engagement for the governance of emerging technologies. *Public Understanding of Science* 23(1): 38–42.

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2015. Global bioethics: What for? http://unesdoc.unesco.org/images/0023/002311/231159e.pdf. Accessed 12 Mar.

United Nations Educational Social Cultural Organisation. 2003. International declaration on human genetic data. http://unesdoc.unesco.org/images/0013/001331/133171e.pdf#page=45. Accessed 5 Sept 2015.

United Nations Educational Social Cultural Organisation. 2005. Universal Declaration on Bioethics and Human Rights. http://unesdoc.unesco.org/images/0014/001461/146180E.pdf. Accessed 27 Feb 2015.

United Nations Office of the High Commissioner for Human Rights. 1998. The Universal Declaration on the Human Genome and Human Rights. http://www.ohchr.org/EN/ProfessionalInterest/Pages/HumanGenomeAndHumanRights.asp. Accessed 22 Apr 2015.

Walker, Peter. 2012. DNA of 100,000 people to be mapped for NHS. *Guardian*, December 10.

Wilsdon, James, and Rebecca Willis. 2004. See through science: Why public engagement needs to move upstream. DEMOS. http://www.demos.co.uk/files/Seethroughsciencefinal.pdf. Accessed 19 May 2015.

Written Evidence Letter from SmithKline Beecham. 2000. Select Committee on Science and Technology. http://www.publications.parliament.uk/pa/ld199900/ldselect/ldsctech/115/115we45.htm. Accessed 25 Mar 2015.

Written Evidence Memorandum by Glaxo Wellcome. 2000. Select Committee on Science and Technology. http://www.publications.parliament.uk/pa/ld199900/ldselect/ldsctech/115/115we24.htm. Accessed 12 May 2015.

Wynne, Brian. 2011. Lab work goes social, and vice versa: Strategising public engagement processes. *Science and Engineering Ethics* 17(4): 791–800.

Younger Committee. 1972. *Report of the committee on privacy, Cmnd. 5012*. London: HMSO.

Part II
Privacy and Data Protection

Many Have It Wrong – Samples Do Contain Personal Data: The Data Protection Regulation as a Superior Framework to Protect Donor Interests in Biobanking and Genomic Research

Dara Hallinan and Paul De Hert

Abstract Genomic research relies on the availability of genomic data. Detached biological samples, stored in facilities known as biobanks, are the source of this data. Donors have interests in these samples. In particular, donors have interests in samples by virtue of the personal data they contain. In relation to this observation, this article puts forward three arguments. *First*: The current European legislative framework relating to samples is inadequate. This inadequacy results from not understanding samples in terms of the information they contain. *Second*: European data protection law, in particular as outlined in the forthcoming Data Protection Regulation, might be looked as a source of solutions. However, whether data protection law can apply to samples at all remains a subject of debate. One key argument supports the position that it cannot: Samples are not data, but rather are physical mater, and therefore can only a source of data. *Third*: The assertion that 'samples are not data, but rather only physical matter' is flawed. Samples do contain data – DNA is data. DNA is understood as information both popularly and in the genetic sciences. In fact, even in informatics, DNA can be understood as data.

1 Introduction

Increasing numbers of biological samples are being collected to be used in genomic research. These samples are extracted from human subjects then stored in facilities known as biobanks. Following storage, samples are subject to a process aimed at

D. Hallinan (✉)
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,
Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen 76344, Germany
e-mail: dara.hallinan@fiz-karlsruhe.de

P. De Hert
Law, Science, Technology and Society (LSTS), Vrije Universiteit Brussel-RC-Juri,
Building B, Room 4B317, Pleinlaan 2, Brussels 1050, Belgium
e-mail: paul.de.hert@uvt.nl

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_6

digitally recording their genetic information – sequencing. At each stage of this process – which consists of four stages (see *below*) – the donor may be seen to have interests. When a sample is extracted (stage 1), donors have interests in how researchers interact with their bodies. When the donor's genetic information is used (stages 3 and 4), donors have interests in how their personal data are handled. In relation to the stored sample (stage 2), however, the donor might be seen to have interests engaged both by virtue of the sample being extracted from their body, and in terms of the personal information that sample contains.

Ideally, the law would recognise the samples in terms of their information content and provide the appropriate protection. However, a brief look at how samples are currently regulated in European Member States reveals a less than satisfactory picture. Only certain countries recognise samples in terms of their informational content and apply data protection rules. Others see samples only in terms of physical matter and accordingly use other regulatory devices to frame and protect samples.

The different legal approaches to regulating samples ends up in a fragmented legal framework which is complicated and even contradictory – good for neither donor, nor researcher. In turn, a persistent 'vagueness' about the applicability of data protection rules creates a legal landscape where the protection of donor interests in personal information contained in samples is often inadequate.

The forthcoming Data Protection Regulation is a legal instrument which has been little considered as a tool for better regulation of biobanking and genomic research. Yet, it seems rather suitable for addressing the problems of law relating to samples. However, before any comprehensive investigation can be conducted into whether and how the Regulation might be of assistance, an initial obstacle must be overcome. It must be possible to show that samples can fall within the scope of application of the Regulation. On the basis of the argument that 'the sample is not data, but is a source of data', it has often been suggested that they cannot.

The aim of this chapter is to argue that samples can be seen as 'data carriers' and that DNA – the 'critical part' of the sample in terms of information content – can certainly be understood in terms of 'personal data'. Accordingly, the article argues that samples can certainly fall within the scope of data protection law.

The genomic research process consists of four steps; collection of sample, storage of sample, sequencing of data, use of sequenced data. Although the collection of the sample may take place locally, samples may then be exchanged and used internationally (Sect. 1). At each phase of this process, the donor has different interests engaged. In relation to stored samples, obtained in the second stage, relevant interests can be understood in two ways – in terms of the sample as physical matter, and in terms of the sample as information (Sect. 2). Following from this observation, the article puts forward three arguments, which build on one another.

First: The current European legislative framework relating to samples is inadequate. This inadequacy results from not understanding samples in terms of information. A number of European states have not understood samples in terms of information and have not regulated accordingly. This is problematic both in terms

of providing adequate protection for donor interests and in terms of the coherence of the legal framework (Sect. 3).

Second: European data protection law, in particular as outlined in the forthcoming Data Protection Regulation, might be looked as a source of solutions. The Regulation offers a harmonized European approach, providing comprehensive protection for interests engaged by the use of personal information. Accordingly, the Regulation might be looked to as a legal instrument which could provide solutions to the discussed problems in sample regulation. However, in order to look at the Regulation as a potential source of solutions, it must be able to apply to samples. One key argument asserts that it cannot: Samples are not data, but rather are physical matter and therefore cannot be 'personal data' and therefore cannot constitute the subject of data protection law (Sects. 4 and 5).

Third: The key argument 'samples are not data, but rather are physical matter' is flawed. Samples do contain data – DNA is data. The Regulation should apply. A USB is physical matter, but data protection law clearly applies to USBs containing personal data. If a sample can be seen to contain data, then surely the sample should be regarded as an 'information carrier', like a USB. DNA is the part of the sample analysed for informational content. Popular and scientific understandings of DNA all recognise it in terms of information. Some who support the 'samples are not data' argument suggest that data protection law relies on concepts of 'data' and 'information' drawn from informatics, and that DNA is excluded by these definitions. Even this claim turns out to be flawed. There is good reason to doubt whether data protection law really relies on informatics definitions. In turn, there is considerable disagreement within informatics as to the meanings of these concepts. Under many proposed definitions, including the most authoritative available – that of the ISO – DNA is data (Sects. 6 and 7).

1.1 The Genomic Research Process: Four Stages

A genome is the complete string of DNA possessed by an organism. Human genomic research aims at understanding how the human genome functions – in particular, how variations in the genome contribute to disease causation. The process of genomic research can be split into four stages; sample collection, sample storage, sequencing of collected samples and finally, the analysis of sequenced genomic data to produce information.

In the first stage, a biological sample, from which the genome can be obtained, will be extracted from the physical body of an individual. This will either result from a specific process, undertaken with the purpose of procuring a biological sample for use in research. For example, a potential donor will be contacted by researchers asking if they would like to participate in genetic research. Alternatively, samples may be sourced from medical procedures which required the removal of human tissue (Riegman et al. 2008, pp. 214–215).

In the second stage, samples are stored – in facilities called biobanks.[1] There are a number of storage techniques available, some of which allow samples to be effectively stored for very long periods of time. For example, modern ultra-low temperature freezers allow storage at −190 °C, a temperature at which biological activity stops and samples can be kept in suspended animation indefinitely (Asslaber and Zatloukal 2007, pp. 195–196).

In the third phase, the biological sample is sequenced to produce genomic information. DNA consists of four organic chemicals – nucleotides denoted by the letters A, C, T and G. A complete human genome consists of a string of over three billion nucleotides. The order and location of nucleotides on the string of DNA defines biological function (Hartl and Ruvolo 2012, p. 7). Sequencing is the process through which the order of nucleotides in a genome is determined, and digitally recorded. The sequencing process is becoming ever faster, cheaper and more automated.[2]

Finally, in the fourth stage, sequenced genomic data is used by researchers to produce information. This information can be general – i.e. not about a specific individual, but rather about human biology in general. Such information is often also referred to as knowledge.[3] Alternatively, a genome can be analysed to produce information specific to an individual – although it may also be possible to extrapolate information about that individual's genetic relatives, or the genetic groups to which the individual belongs.[4]

The first stage of the genomic research process is geographically specific. Indeed, it must even take place within one specific institution. Collection is a physical act which must take place in a specific location. However, following extraction, there

[1] For a definition see National Health and Medical Research Council (2010).

[2] The first full genome was sequenced by the Human Genome Project in 2001. The sequencing process took 10 years, required a $3bn investment and was the product of the collaborative efforts of 200 scientists from institutions located all over the globe (International Human Genome Consortium et al. 2001, pp. 860–921). By comparison, in 2014, Illumina, a manufacturer of genetic sequencing equipment, brought out the Illumina Hi Seq X-10 sequencing machine. According to Illumina's product description, the High Seq X-10 is capable of sequencing 49 genomes per day at a cost of only $1000 each (Illumina 2015).

[3] For example, sequenced genomic data may be used in Genome Wide Association Survey research. In such research, thousands of genomes from individuals displaying a certain trait are compared with thousands of other genomes from individuals not displaying this trait. The comparison of genomes allows the production of generalized information about significant points of difference between the two sets of genomes. These points of difference hold information as to the genetic basis for the studied trait. See Kaye (2012, pp. 36–38).

[4] If knowledge is available as to the significance of a certain type of genetic architecture, the detection of that architecture in a specific genome can reveal information about the person from whom the genome came. Certain of the genetic characteristics observed may be biographically highly significant – for example, those relating to disease predisposition. See Hallinan and De Hert (2015).

is an increasing trend toward making samples and data available for researchers around Europe, and even internationally.[5]

The possibility to exchange samples is facilitated and expedited by a general trend toward the networking of biobanks – both nationally and internationally (Asslaber and Zatloukal 2007, pp. 196–201). In many such endeavours, a set of rules for collaboration are agreed upon in advance and a central hub functions as a virtual biobank providing a catalogue – often online – of materials available across the network. This catalogue is available to be searched through by researchers in need of specific types of samples. In Europe, a number of national networks exist – the Telethon network of biobanks consists only of Italian Biobanks.[6] Within the last few years, however, there has also been considerable progression towards European level biobanking networks.[7]

2 Donor Interests Everywhere: Even with Regard to Detached Samples (Stage 2)

At each point in the genomic research process, different materials are being handled, and different acts are being conducted. The whole process, described *above*, can be seen in terms of the conversion of the corporeal into the informational. Accordingly, donor interests will be different at different points in the process. Which donor interests are seen as relevant at any point in the process, will thus depend on the

[5] Asslaber observes that 'in a genome scan for a genetic polymorphism associated with a certain disease, DNA of about 10,000 diseased individuals should be analysed' (Asslaber and Zatloukal 2007, p. 194). For any single biobank to collect this number of samples of individuals displaying the relevant form of the disease will be a long and arduous process – especially for rare diseases. Nevertheless, reaching a critical mass of samples can be significantly expedited 'if biobanks cooperate . . . so that cases from different biobanks can be combined' (Asslaber and Zatloukal 2007, p. 194). Further, certain approaches to research will require the availability of specific types of samples for which certain genetic or environmental variables have been removed – for example, samples of a particularly homogenous ethnic population. Identifying relevant biomarkers becomes much simpler when variation can be minimized. Availability of such samples may be geographically specific, although research may take place globally. To facilitate such research, biobanks need to exchange samples and collaborate across borders.

[6] Telethon Network of Genetic Biobanks. http://www.biobanknetwork.org/members.php. Accessed 03 July 2015.

[7] Perhaps the two most prominently discussed networks are the EuroBioBank network for scientists studying rare diseases and the Biobanking and BioMolecular resources Research Infrastructure (BBMRI) (Eurobiobank. http://www.eurobiobank.org/en/partners/partners.htm Accessed 03 July 2015; BBMRI. http://bbmri-eric.eu/memberstates. Accessed 03 July 2015). BBMRI currently consists of over 280 organisations and is particularly interesting as it represents a directed effort by the European Commission to create a European biobanking network. Finally, there are global networks, such as the Public Population Project in Genomics (P3G) which boasts truly global membership (Public Population Project in Genomics. http://p3g.org/membership/institutional-members. Accessed 03 July 2015).

perception of which type of substance – corporeal or informational – is being acted upon. At the first stage of the process, tissue is being extracted from the body of the donor. Donor interests will thus relate to when researchers may engage with the physical body. At stages three and four, data or information are being processed. The type of donor interests engaged will thus relate to the use of personal information. In these stages, there is relatively little dispute as to how the material under consideration should be perceived. However, at the second stage of the process – in relation to detached, but unsequenced, samples – the type of substance being acted upon might be conceived of in two ways. This has significant consequences for the types of donor interest which might be regarded as relevant.

First, the detached sample might be seen in terms of physical bodily material. Traditional interests related to the body are hard to assert as there is no longer any interaction with the living body. However, the donor might still be seen to have a relationship with the detached material by virtue of its extraction from their body.

The recognition of such an interest in bodily material can be found in legislation applicable to the genomic research process. For example, the UK's Human Tissue Act (UK Parliament 2004a). The genesis of the Act was the outcry that followed the Bristol Royal Infirmary and Alder Hey scandals. In these cases, it came to light that retention of bodily material from dead children – including organs, foetuses and stillborn children – had taken place with no, or inadequate, consent. The Act was an attempt to balance 'the rights and expectations of individuals [in controlling access to their tissue] . . . and broader considerations such as research' (UK Parliament 2004b).

Second, and in our opinion more importantly, the detached sample might be seen in terms of information. Genetic samples are only collected so that the information they contain may be processed and analysed. They are searched for and exchanged by researchers on the basis of the information they are presumed to contain. In turn any party collecting samples for genetic research will also have the means available to them to convert the sample into digital data. As the sequencing process has become faster and cheaper, the distance between sample and eventual digital data shrinks accordingly. Anything that can be done with sequenced genetic information, can also be done with the original sample and a sequencing machine. Donor interests in the use of their personal information might be seen to be founded on two recognitions. If the donor has interests in controlling who can use their personal data, and for which purposes, it is fair to state that the donor should have those same interests relating to who has access to their samples and what they are used for (Bygrave 2010, p. 1).

The interpretation of genetic samples as substances capable of engaging informational privacy interests has received academic as well as legal recognition in Europe (Taylor 2012, pp. 161–165). For example, the European Court of Human Rights followed this line of reasoning in the *Marper* case when it stated: 'The Court notes that . . . cellular samples . . . constitute personal data' (European Court of Human Rights 2008, §68).

Given that the storage and use of samples in genomic research can be conceived of in terms of information, we might now consider how this fact has been translated

into legal protection. Two questions are particularly relevant. First: Does the law perceive samples, and the handling of samples, in terms of information? Second: If not, which problems does this cause?

3 Problems Arising from Legal Systems Approaching Samples Without Data Protection

The legal framework around the storage and exchange of samples in Europe has, as yet, not been harmonized. As a result, Member States have taken their own approaches. These differ considerably.

On the one hand, there are European states with a data protection approach to samples: samples are regulated under the same laws applicable to sequenced data – for example, Estonia (Estonian Parliament 2000, §7(1)). These states can be seen to regard samples in terms of information and to offer legal protection based on data protection rules. Accordingly, donors benefit from a relatively well developed – albeit perhaps imperfect – system of protection for donor interests in personal information (see *below*).

On the other hand, there are European states which have taken alternative approaches, and which regard the sample in terms of physical matter. These states regulate detached samples without turning to data protection rules. Although the perception of the regulated object might be comparable between these states, the approaches taken to regulation may still differ considerably. For example, England regulates sample collection and use under the Human Tissue Act – an act designed primarily to protect a *sui generis* interest in tissue detached from the body (UK UK Parliament 2004a). This can be compared with the approach taken by Germany, in which a property law approach is favoured – the detached sample is seen as the donor's property (Albers 2013, p. 486).

There are three problems with the current framework in Europe. One related to the substantive protection offered to donors interests in their personal information. Two related to the legal fragmentation caused by differing approaches. These problems particularly challenge those Member States that work with a regulatory approach *not* based on data protection rules.

Firstly, when legislation is not built on the recognition of the informational properties of samples, this tends to leave gaps in the protection of interests related to the processing of personal data. When tissue is understood in terms of information, donor interests are relevant whenever tissue is handled which contains an individual's genome. When protection is engaged on the basis of the connection between the donor and their physical sample, donor interests only exist if that connection is still present. The scope of approaches based on the connection between the donor and their 'physical material' can thus be narrower than those based on the informational content of samples. For example, in the UK, the genesis of the Human Tissue Act was to protect donor interests in detached samples by virtue of donors having an ongoing relationship with their bodily material (UK Parliament 2004b).

Accordingly, as soon as the connection between donor and material no longer exists, the Act ceases to apply (Human Tissue Authority 2014a). It is now possible to make copies of cellular material with the same genetic information contained within – for example, through the creation of an immortal cell-line. Such new material is not regarded as 'relevant material' under the act as it is not 'manufactured' in the body of the donor (Human Tissue Authority 2014a). Despite the same information being contained within the cells, the connection between donor and material is judged no longer to exist, the Act no longer applies and the donor enjoys absolutely no more protection.

In turn, the range of protection offered to donors under 'physical material' approaches may be reductive. Data protection law – the law protecting individuals' interests engaged by the processing of their personal data – outlines a broad set of rights, applicable before, during and after processing (see next Sect. 5). Approaches which recognise the sample in terms of 'physical material' can be very limited in the rights they give to the donor. For example, under the German approach, as soon as property rights in the sample are signed away, it is highly uncertain which relationship, if any, the donor is seen to retain with the sample (Albers 2013, pp. 486–487).

Secondly, the genomic research infrastructure is increasingly set up as an international endeavour. In regulating an international endeavour, a harmonized legal framework is preferable to a fragmented framework. Currently, the differing approaches taken by European states means that different sets of obligations need to be followed whenever samples are to be transferred across borders. The fact that there are different approaches stems from the fact that samples are perceived differently by different states. The need to follow these differing obligations has been seen to make the transfer of samples between European countries bureaucratic and complicated (Gibbons 2012, p. 90). It is hard for researchers to distil the relevant requirements and even harder to work out how they might be met. Eventually, this has two negative consequences: the confusion and red tape created can hinder the smooth function of research *and*, in cases of international transfer, it can be hard for donor subjects to understand which protections they are entitled to and when these apply.

Finally, in genomic research, samples and data are increasingly proximate. Their proximity increases with developments in sequencing technology. It is now only a very small step between possession of a sample and the production of sequenced data. In turn, they are each part of the same process serving the ends of scientific research. In this regard, it has been observed that they essentially travel together. Accordingly, it makes sense that any obligations relating to samples and data should be as streamlined as possible. However, when samples and data are treated as legally distinct entities, each is subject to its own system of protection. In the best case, regimes coincide. However, there are also cases in which there are differences, or even contradictions, between the two systems. For example, in the UK, the Human Tissue Act lays out a set of rules for research subject consent. These rules allow the research subject to give broad consent to the use of their bodily materials – i.e. that materials can be used in any future research project – in biobank research (Human

Tissue Authority 2014b, p. 11). However, as soon as data are sequenced from the sample, the main piece of applicable legislation is the Data Protection Act. It is far from clear that broad consent is a legitimate form of consent in data protection law, which requires consent to be specific to a processing operation (Hallinan and Friedewald 2015, pp. 10–19). It is useless to have a broad consent in relation to the possible uses of the sample in research when no more than specific data can be generated and used. Once again, the need to follow different forms of rules for sample and data is bureaucratic and confusing. This serves neither research, nor donor interests.

Give that the current situation is unsatisfactory, we might look to other areas of law for possible solutions. One area of law which may be highly useful in the regulation of genomic research, but whose potential in this regard has still been relatively little considered, is data protection law.

4 A Short Guide to Data Protection Law in the EU and the Data Protection Regulation

Data protection law is the area of law outlining when personal data may, and may not, be processed. Data protection law in Europe has twin goals: (1) to ensure the free flow of data through a harmonized set of laws; and (2) to provide a high standard of protection for fundamental rights – in particular informational privacy – which could be impacted by the processing of personal data.

The current piece of legislation defining European data protection law is Directive 95/46 (European Parliament and European Council 1995). However, over time, its relevance and suitability have come into question. First, Directives need transposition into national law. The differences between national transpositions of the Directive are significant and accordingly, it has been seen to have failed to adequately harmonize legislation. Second, there has been considerable technological change in the 20 years since the Directive was drafted. The Directive has been argued to be inadequately equipped to effectively deal with the opportunities and risks in data processing which have come with this change (European Commission 2010, pp 1–4).

Accordingly, 3 years ago, a process of reform of data protection law was started and in January 2012, the Commission released the proposed Data Protection Regulation, offered as a replacement to Directive 95/46 (European Commission 2012). In October of 2014 the Regulation (with certain amendments) was successfully voted on before the European Parliament. In June 2015, the Council released their common position on the reform. The reform is now subject to trilogue negotiations before finally being made law.[8]

[8] All three text drafts thus far available and compared at: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

Unlike the Directive – which required national transpositions – the Regulation has direct affect and will be legally binding in all European states. This will serve to address the fragmentation which resulted from differing transpositions of the Directive. In turn, the Regulation has been drafted complete with a binding European level interpretation mechanism.[9] This allows the Regulation's mechanisms to be adapted and interpreted, at the European level, to deal with further technological change.

The Regulation's scope extends to processing done in a wide range of contexts and by a wide range of actors whenever the 'personal data' of a 'data subject' is processed. There was no mention of genetic data in the text of Directive 95/46, and this led to some uncertainty as to the Directive's applicability. However, the Regulation has clarified this by specifically listing genetic data as a category of 'sensitive data' under Article 9.[10] Given that genetic data cannot be anonymised, it is the case that genetic data should always be regarded as 'personal data'. Whenever genetic data are processed, the Regulation will apply (Hallinan et al. 2013, pp. 317–329).

The Regulation provides a comprehensive approach to protecting informational privacy interests throughout the lifetime of a data processing operation. This approach can be subdivided into four core mechanisms.

1. Proposed processing must be checked in advance for proportionality by the supervisory authority.
2. Whenever personal data are processed, the data controller is subject to rules as to when and how they may be processed. This approach lays out a set of procedural, technical and organisational aspects of data processing – including rules outlining fair information practises.
3. Where the data controller does not have an overriding interest in the processing of data, the data subject is granted control over whether data may be processed and must be asked for consent. Even where the data controller has the right to process without consent, the controller must make processing transparent to the data subject and the data subject retains certain rights relating to their data.
4. The Regulation provides for independent oversight to make sure that processing continues in a legitimate fashion. When processing is found to be illegitimate, or disproportionate, the Regulation lays out possibilities for redress (Beyleveld 2004, pp. 8–21).

At first glance then, there is good reason to look closer at whether the Regulation might be able to do some work in remedying issues with the current legal framework applicable to samples. If the Regulation could apply to samples, this would provide comprehensive protection for donor interests in personal information. Equally, if the Regulation could apply to samples, this might do some work toward addressing the issues of legal fragmentation. The Regulation would apply in a harmonized way

[9]See Chapter VI in each version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

[10]See Article 9 in each version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

around Europe, as well as outlining a harmonized set of requirements applicable to the use of samples and sequenced data.

However, before the Regulation can be touted as any kind of resolution, a number of more subtle questions must be asked. For example: Could the Regulation simply replace all diverging approaches, or are there legislative functions served by the various approaches which the Regulation could not replicate? Will the definitions and mechanisms of the Regulation function in relation to samples? However, before asking any such questions, one key question must be clarified: Can the Regulation even apply to samples?

5 An Unanswered Question: Will the Regulation Apply to Samples?

All legislation has a scope of application – a delineation of what it aims to regulate. The Regulation could only be used to apply to samples, should its scope be adequately broad to extend to samples. In essence, it must be clarified that samples constitute the *rationae materiae* of the Regulation. Initially, we might look at the text of the Regulation for answers. However, the Articles defining the scope of the Regulation remains silent on this point. In turn, authoritative legal sources offer conflicting interpretations.

In the text of the Regulation, two Articles are particularly important in defining scope – Article 2(1) and article 4(1).

The goal of Article 2(1) is to delineate scope. The Article states: 'the processing of personal data wholly or partly by automated means, and to the processing other than by automated means of data which form part of a filing system or are intended to form part of a filing system'.[11] Genetic samples stored for research are organised and searchable in a way very similar to any other form of filing system. Accordingly, the fact that no automatic, or computer-based, processing has taken place at the storage phase, is not an obstruction to the application of the Regulation. However, Article 2(1) not only requires a filing system to be present, but also that 'personal data' are contained in this filing system.

'Personal data' are defined in Article 4(1). Article 4(1) states that personal data are defined as: 'any information relating to an identified or identifiable natural person ('data subject')'.[12] There is relatively little difficulty in conceiving genetic samples as being capable of relating to an identified or identifiable natural person. Such a relationship might follow from their being collected alongside identifying information (Sweeny et al. 2013). A relationship may also follow from DNA

[11] See Article 2(1) or equivalent in each version: See Chapter VI in each version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

[12] See Article 4(1) or 4(2) in each version (there are minor language differences between versions, but these do not change the applicability of the analysis): See Chapter VI in each version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

matching procedures aimed at identifying a specific individual (Nuffield Council on Bioethics 2007, pp. 8–11). However, the definitions provided offer no further help in clarifying what is meant by the terms 'data' or 'information' – which seem to be used as synonyms in Article 4(1). Without a further definition of what is understood under these terms, it is impossible to conclude whether samples can be conceived of as 'personal data' or not.

Ordinarily, when a legal text does not sufficiently answer a question, one can look to other authoritative sources to provide clarity as to how the law should be interpreted. Unfortunately, in this case, authoritative sources conflict.

On the one hand, there are authoritative sources which suggest that the Regulation could be applied to genetic samples. First, the intention behind data protection law indicates the concepts and definitions of the law should be given a broad interpretation. The intention is to provide protection for data subject interests whenever others are interacting with their personal information. Data protection law, the Regulation included, was thus always intended to be flexible to adapt to deal with new technological possibilities in data processing (European Commission 2012, p. 12). This thinking is evidenced in original guidance as to the concept of 'personal data'. The Commission stated that 'personal data' be interpreted 'as general[ly] as possible, so as to include all information concerning an identifiable individual' (European Commission 1992, p. 10). In interacting with samples, researchers are doing the equivalent of interacting with an individual's personal data. Accordingly, the storage and use of genetic samples as sources of data might simply be seen as yet another novel development in data processing. If a broad interpretation of 'personal data' is desired, then perhaps samples should be covered within the scope of the Regulation. Second, authoritative legal fora have recognised that genetic samples can constitute 'personal data'. For example, the European Court of Human Rights categorically stated in the *Marper* case that samples should be regarded as data (European Court of Human Rights 2008, §68).

On the other hand, however, other authoritative interpretations suggest that samples should not be regarded as 'personal data'. First, if samples are to be regarded as data at all, then surely they would be regarded as genetic data. In the Council's version of the Regulation – the latest version to be released – Recital 25(a) elaborates the scope of the term 'genetic data'. It states that 'genetic data' should only be regarded as: 'resulting from an analysis of a biological sample from the individual in question'. Such a definition might be read as displaying an intention to exclude genetic samples from falling within the scope of application of the Regulation.[13] Second, the question has been explicitly considered by the Article 29 Working Party in relation to the Directive. The Article 29 Working Party is specifically tasked with providing authoritative interpretations of European data protection law. They state: 'Human tissue samples (like a blood sample) are themselves sources out of which biometric data are extracted, but they are not

[13] See Recital 25(a) in the Council version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

biometric data themselves (as for instance a pattern for fingerprints is biometric data, but the finger itself is not). Therefore the extraction of information from the samples is collection of personal data, to which the rules of the Directive apply. The collection, storage and use of tissue samples themselves may be subject to separate sets of rules' (Article 29 Data Protection Working Party 2007, p. 9).

This last objection outlined by the Article 29 Working Party – that samples are not data, but are physical matter and therefore cannot be 'personal data' – is particularly significant. In fact, each time it is argued – in jurisprudence or academia – that data protection law cannot apply to samples, this argument takes centre stage (Beyleveld et al. 2004, p. 428; Government of Australia 2005, pp. 8–9). However, it is notable how seldom this objection is further elaborated. For example, the Article 29 Working Party do not actually elaborate what they mean by 'data' or 'information'. In turn, they do not elaborate exactly why samples cannot be understood to be included under these terms. Perhaps this lack of clarification sits on the presumption that 'samples cannot be data' is a self-explanatory statement?

We do not see this as self-explanatory. To explain why, we might rephrase the objection. A USB stick can be undoubtedly described as solid matter. However, a USB stick containing personal data is also undoubtedly the subject of data protection law insofar as that USB stick is used in the processing of personal data.[14] This is as a result of the USB stick 'containing' personal data. If the sample can be seen to contain data, in the same way as the USB stick might be seen to contain data, then should data protection law not also apply to samples (Albers 2013, p. 487)? To demonstrate how samples can be seen to contain data, it is thus necessary to more closely examine the 'critical parts' of the sample. The most obvious unit of analysis in this regard is DNA.[15]

6 DNA as Information: The Dominant Understanding of DNA in Genetic Science

It is not difficult to find references to DNA in terms of data, or information. In fact, the dominant popular metaphors for understanding DNA are almost all informational – 'a genetic code', the genome as a 'book' or 'a blueprint'.[16] Each of these metaphors liken DNA to a medium through which information is transferred.

[14] See Article 2 of each version: http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf. Accessed 03 July 2015.

[15] The one place this idea seems to have been given more extensive legal consideration is; in the Australian Law Reform Commission and Australian Health Ethics Committee, *'Essentially Yours: The Protection of Human Genetic Information in Australia'* report (2003, p. 268). Unfortunately, the analogy is only partially outlined, and its significance is not carried forward in further analysis.

[16] See, for example, the numerous information metaphors in US President Clinton's; 'Remarks made by the President ... on the Completion of the First Survey of the Entire Human Genome Project' (2000).

However, the idea of DNA as information is not just a useful popular metaphor for understanding genetics. In fact, the root of such metaphors is to be found in genetic science itself.

For over 50 years, starting shortly after Watson and Crick elaborated the structure of the DNA molecule – the famous double helix – the genetic sciences have relied heavily on information theory to understand the structure and function of DNA (Griffiths and Stotz 2013, pp. 143–153).

Using information theory, geneticists began to understand DNA no longer only as a physical molecule, but much more as the medium through which biological information is transferred. The language in which this biological information is encoded, consists of four letters – the nucleotides A, C, T and G. The grammar for this language he is provided by the order in which nucleotides are laid out on in the 3.2 billion long nucleotide chain. The type of information which is transferred is that of biological specificity – i.e., how an organism might go about producing certain physical states. The sender of the information are the parent organisms. The recipient of the information is the offspring organism, to whom the genome belongs.

In some contexts, the extent to which use of information theory in understanding DNA has value is still debated. For example, there are questions as to the value of describing human development, or socialisation, in terms of the result of information contained in DNA – as the result of the genetic code. However, that there is a 'genetic code' and that this is contained in DNA is now virtually unchallenged (Griffiths and Stotz 2013, pp. 153–158). The understanding of the DNA molecule in terms of information is thus unproblematic from the perspective of genetic science. Prominent evolutionary biologist George Williams even went so far as to say that it makes more sense to conceive of genes as units of information, rather than physical objects made of DNA (Willimas 1992).

Indeed, so strong are the parallels between DNA and information, that DNA has been proposed as an alternative to digital storage media. George Church at Harvard University, for example, has succeeded in recording copies of his latest book in a strand of DNA (Anthony 2012). Church observes that 'as digital information continues to accumulate, higher density and longer term storage solutions and necessary. DNA has many potential advantages as a medium for immutable, high latency information storage needs' (Church et al. 2012). If physical samples cannot be data or information, then does George Church's DNA book not constitute data or information either?

However, despite the fact that DNA are regarded in terms of information in genetic science, this is not definitive proof that the 'samples cannot be data' objection is flawed as it applies to data protection law and the Regulation. The objection relates to the scope of concepts in data protection law. Genetic science is not the source of conclusive proof as to the meaning of concepts in data protection law.

In this regard, it has been suggested that data protection law has been built on a concept of 'data' and 'information' drawn from informatics – the science of computer information systems (Bygrave 2010, p. 14; Nys 2004, p. 41). Although this is not specifically clarified in any policy document, it seems a reasonable assertion,

not least as data protection law developed as a response to automated processing by computers. Is it then the case that, using informatics definitions of 'data' and 'information', DNA is precluded from qualifying as 'data' or 'information'?

7 DNA Can Be Understood as 'Data' in Informatics

Two arguments can be put forward against the idea that an informatics concept of data can be relied upon to exclude samples from being within the scope of the Regulation. The first relates to whether the Regulation really sits on concepts of 'data' and 'information' drawn from informatics. The second challenges the presumption that informatics concepts of 'data' and 'information' cannot subsume DNA.

There is a body of literature in informatics devoted to the discussion of the meaning of the concepts of data and information in informatics. As a start point for our analysis, we have relied heavily on Chaim Zins 2007 work: 'Conceptual approaches for defining data, information and knowledge [in informatics]' (Zins 2007). In this work, Zins documented definitions for these terms provided by 45 different informatics scholars.

First, if this statement is true: 'data protection law, and by extension the Regulation are implicitly built on concepts of 'data' and 'information' drawn from informatics', then it would be fair to assume that informatics and data protection uses of the terms should not conflict.

It seems that this is not always the case. For example, in informatics, there are significant differences between the concepts of 'data' and 'information' (Zins 2007, p. 479). In data protection law to date, the terms have largely been used interchangeably. Indeed, certain data protection scholars have specifically noted the lack of differentiation between 'data' and 'information' understood in an informatics sense, as a problem for data protection law (Manson 2009, pp. 26–28). Such an observation serves to cast doubt on the claim that informatics concepts can be looked to as authorities in the definition of the same terms in data protection law. Should informatics have been the source for concepts in data protection law, surely such a significant difference would not have been overlooked.

Second, if this statement is true: 'informatics concepts of 'data' and 'information' preclude inclusion of DNA', then it might be fair to assume: (1) There are relatively clearly defined concepts of 'data' and 'information' in informatics. (2) That these clearly exclude DNA from their scope. Both assumptions are open to question.

There are considerable disagreements between informatics scholars as to the meanings of information and data. In Zins work, on a number of key points, scholars definitions diverged. Differences related to the scope of the concepts, to qualifying characteristics necessary to be considered as either 'information' or 'data' and to the differences between concepts (Zins 2007, pp. 487–489). In considering whether DNA can be understood as either 'data' or 'information', these points of difference can be significant.

On the one hand, definitions provided by certain scholars would appear to exclude DNA. Such definitions tended to focus on; the need for machine involvement in creation or storage, conscious human involvement in collection or on the need for representation in alphabetic or numeric form, of 'data' or 'information'. For example, scholars who can be seen to fall into this category state: 'Data are formalized parts (i.e. digitalized contents) of socio-cultural information', or 'Data are the raw observations about the world collected by scientists and others, with a minimum of contextual interpretation' (Zins 2007, p. 484).

On the other hand, however, other definitions were offered under which DNA could certainly be included. For example: 'Data are perceptible or perceived—if and when the signal can be interpreted by the 'user'—attributes of physical, biological, social or conceptual entities' or 'Data are a representation of facts or ideas in a formalized manner, and hence capable of being communicated or manipulated by some process' (Zins 2007, pp. 485–486).

Such an inclusive conceptualisation of data and information is shared by the most authoritative source outlining the meaning of these terms in informatics: ISO 2382–1, 'Information technology – Vocabulary – Part 1: Fundamental terms' (International Standards Organisation 1993). In this document, the International Standards Organization defined data as: A reinterpretable representation of *information* in a formalized manner suitable for communication, interpretation, or processing . . . Data can be processed by humans or by *automatic* means' (International Standards Organisation 1993). They retain this definition in their 2015 revision. This definition can be broken down to clearly show how DNA could fall within its scope. DNA are reinterpretable – otherwise they would be useless, both as a means to transfer biological specificity between generations, as well as for all forms of genetic analysis. DNA is a representation of information – information as to biological specificity. DNA is 'written' in a formalized language – comprised of four nucleotides. Finally, through the sequencing and analysis process, DNA is clearly capable of being processed by human and automatic means.

Accordingly, the argument which suggests that an informatics definition sits behind the concept of 'personal data' in data protection law, and that DNA cannot be subsumed within such an informatics definition, can be challenged.

As a result, we believe it is possible for genetic samples to fall under the scope of the Regulation without encountering any conceptual problem.

8 Conclusion

The collection of human samples in biobanks raises a number of issues, amongst which is the protection of donor interests is particularly important. In particular, donors have an interest in samples by virtue of their information content.

We have argued that the current European legal framework for the regulation of samples is inadequate. Only a few states recognise samples in terms of their information content. On the one hand, this means donor interests in the use of

their personal information are not always adequately protected. On the other hand, the different approaches taken to sample regulation leads to a highly fragmented framework.

In turn, we argued that the Data Protection Regulation might be looked to as a solution to some of these problems. It is legislation harmonized at European level, which aims to provide comprehensive protection for individual interests in personal information. However, the application of data protection law to physical samples remains the subject of debate. The text of the Regulation remains silent on the issue. Authoritative sources disagree. However, those sources which question the application of data protection law to samples generally do so on the basis of one objection – that samples are physical matter and therefore cannot be data.

We finally argue that the objection – 'samples are physical matter and therefore cannot be data' – is flawed. We argue that samples contain data – in DNA – and that therefore, the Regulation should apply. It would be absurd to suggest that handling a USB stick containing personal data was not subject to data protection law. But the USB stick is also physical matter. The key issue is that the USB stick contains 'personal data'. DNA is the 'critical part' of the sample as far as information extraction is concerned. If DNA can be regarded in terms of information, then surely the sample can be regarded as an information carrier in a similar way to a USB stick? A cursory search will show that DNA is often referred to in terms of information. In fact, almost all dominant popular conceptualisations of DNA revolve around information. However, such informational conceptualisations of DNA are not limited to popular metaphors. In fact, DNA is also understood in genetic science as information. Although a counter-argument might be put forward – that data protection law relies on concepts of 'data' and 'information' drawn from informatics, and that DNA is excluded by these definitions – this too is weak. First, the claim that data protection law relies on such informatics definitions is not supported in policy documents. In turn, there appear to be contradictions between how the terms 'data' and 'information' are used in data protection law and in the Regulation, and in informatics. Second, there is considerable disagreement within informatics as to the meanings of these concepts. Under many proposed definitions, including the most authoritative we could find – that of the ISO – DNA is data.

Moving forward, a number of questions remain open. First, how the Regulation – its definitions and mechanisms – should best apply to samples needs further research. Second, the foregoing is not to say that other regulatory approaches are deprived of meaning because the Regulation can apply and may offer solutions. On the contrary, data protection has it limits. Data protection is unlikely to be able to fulfil all relevant legislative tasks relating to samples – data protection cannot subsume all aspects of a property approach, for example. In turn, suitable procedural and substantive solutions with regard to samples will not always be immediately follow from the applications of data protection rules. Complementary ideas from other areas of law can help in tracing necessary normative lines and in filling out proportionate procedural approaches (Gutwirth and De Hert 2006). Investigation of these questions will follow in our future work.

References

Albers, Marion. 2013. Rechtsrahmen und Rechtsprobleme bei Biobanken. *Medizinrecht* 31(8): 483–491.

Anthony, Sebastian. 2012. Harvard cracks DNA storage, crams 700 terabytes of data into a single gram. Extremetech, August 17. http://www.extremetech.com/extreme/134672-harvard-cracks-dna-storage-crams-700-terabytes-of-data-into-a-single-gram. Accessed 03 July 2015.

Article 29 Data Protection Working Party. 2007. Opinion 4/2007 on the concept of personal data. WP136.

Asslaber, Martin, and Kurt Zatloukal. 2007. Biobanks: Transnational, European and global networks. *Briefings in Functional Genomics and Proteomics* 6(3): 193–201.

Australian Law Reform Commission and Australian Health Ethics Committee. 2003. Essentially yours: The protection of human genetic information in Australia. http://www.alrc.gov.au/sites/default/files/pdfs/publications/ALRC96_vol2.pdf. Accessed 03 July 2015.

Beyleveld, Deryck. 2004. An overview of directive 95/46/EC in relation to medical research. In *The data protection directive and medical research across Europe*, ed. Deryck Beyleveld et al., 8–21. Aldershot: Ashgate.

Beyleveld, Deryck, Andrew Grubb, David Townend, Ryan Morgan, and Jessica Wright. 2004. The UK's implementation of directive 95/46/EC. In *Implementation of the data protection directive in relation to medical research in Europe*, ed. Deryck Beyleveld et al., 403–428. Aldershot: Ashgate.

Bygrave, Lee. 2010. The body as data? Biobank regulation via the 'Back Door' of data protection law. *Law, Innovation and Technology* 2(1): 1–25.

Church, George, Yuan Gao and Sriram Kosuri. 2012. Next generation digital information storage in DNA. *Science Express*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.299.5153&rep=rep1&type=pdf. Accessed 03 July 2015.

Clinton, William. 2000. Remarks made by the president . . . on the completion of the first survey of the entire human genome project. The White House Office of the Press Secretary. June 26. https://www.genome.gov/10001356. Accessed 03 July 2015.

Estonian Parliament. 2000. Human Genes Research Act. English translation available at: https://www.riigiteataja.ee/en/eli/531102013003/consolide. Accessed 03 July 2015.

European Commission. 1992. Amended proposal for a Council Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data. COM (92) 422 final. http://aei.pitt.edu/10375/1/10375.pdf. Accessed 03 July 2015.

European Commission. 2010. A comprehensive approach on personal data protection in the European Union, COM (2010) 609 final. http://ec.europa.eu/justice/news/consulting_public/0006/com_2010_609_en.pdf. Accessed 03 July 2015.

European Commission. 2012. Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). COM (2012) 11 final. http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf. Accessed 03 July 2015.

European Court of Human Rights. 2008. *S. and Marper v United Kingdom*, no. 30562/04 and 30566/04.

European Parliament and European Council. 1995. On the Protection of individuals with regard to the processing of personal data and on the free movement of such data. Directive 95/46/EC. http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&from=EN. Accessed 03 July 2015.

Government of Australia. 2005. Australian Law Reform Commission and Australian Health Ethics Committee Report essentially yours: The protection of human genetic information in Australia: Government response to recommendations. http://apo.org.au/files/Resource/3newfinalresponse6december2005.pdf. Accessed 03 July 2015.

Gibbons, Susan. 2012. Mapping the regulatory space. In *Governing biobanks: Understanding the interplay between law and practice*, ed. Jane Kaye et al., 51–92. Oxford: Hart Publishing.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and philosophy: An introduction*. Cambridge: Cambridge University Press.

Gutwirth, Serge, and Paul De Hert. 2006. Privacy, data protection and law enforcement. Opacity of the individual and transparency of power. In *Privacy and the criminal law*, ed. E. Claes, A. Duff, and S. Gutwirth, 61–104. Antwerp: Intersentia.

Hallinan, Dara and Michael Friedewald. 2015. Open consent, biobanking and data protection law: Can open consent be 'informed' under the forthcoming data protection regulation? *Life Sciences, Society and Policy* 11(1). doi: 10.1186/s40504-014-0020-9. See http://lsspjournal.springeropen.com/

Hallinan, Dara, Michael Friedewald, and Paul De Hert. 2013. Genetic data and the data protection regulation: Anonymity, multiple subjects and a prohibitionary logic regarding genetic data. *Computer Law & Security Review* 29(4): 317–329.

Hartl, Daniel, and Maryellen Ruvolo. 2012. *Genetics: Analysis of genes and genomes*, 8th ed. Burlington: Jones and Bartlett Publishing.

Human Tissue Authority. 2014a. List of materials considered to be 'relevant material' under the Human Tissue Act 2004. https://www.hta.gov.uk/policies/list-materials-considered-be-%E2%80%98relevant-material%E2%80%99-under-human-tissue-act-2004. Accessed 03 July 2015.

Human Tissue Authority. 2014b. Code of practise 9: Research. https://www.hta.gov.uk/sites/default/files/Code_of_practice_9_-_Research.pdf. Accessed 03 July 2015.

Illumina. 2015. HiSeq X ten specification sheet. http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf. Accessed 03 July 2015.

International Human Genome Consortium et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

International Standards Organisation. 1993. Information technology—Vocabulary—Part 1: Fundamental terms. ISO 2382–1. Revised by ISO/IEC 2382–1, 2015. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=63598. Accessed 03 July 2015.

Kaye, Jane. 2012. Embedding biobanks in a changing context. In *Governing biobanks: Understanding the interplay between Law and practice*, ed. Kaye Jane et al., 30–51. Oxford: Hart Publishing.

Manson, Neil. 2009. The medium and the message: Tissue samples, genetic information and data protection legislation. In *The governance of genetic information: Who decides?* ed. Widdows Heather and Mullen Caroline, 15–37. Cambridge: Cambridge University Press.

National Health and Medical Research Council. 2010. Biobanks information paper. https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e110_biobanks_information_paper_140520.pdf. Accessed 03 July 2015.

Nys, Herman. 2004. Report on the implementation of directive 95/46/EC in Belgian law. In *Implementation of the data protection directive in relation to medical research in Europe*, ed. Deryck Beyleveld et al., 29–41. Aldershot: Ashgate.

Nuffield Council on Bioethics. 2007. The forensic use of bioinformation: ethical issues. http://nuffieldbioethics.org/wp-content/uploads/The-forensic-use-of-bioinformation-ethical-issues.pdf. Accessed 03 July 2015.

Riegman, Peter, et al. 2008. Biobanking for better healthcare. *Molecular Oncology* 2: 213–222.

Sweeny, Latanya, Akua Abu and Julia Winn. 2013. Identifying participants in the personal genome project by name. http://dx.doi.org/10.2139/ssrn.2257732. Accessed 03 July 2015.

Taylor, Mark. 2012. *Genetic data and the law: A critical perspective on privacy protection*. Cambridge: Cambridge University Press.

UK Parliament. 2004a. Human Tissue Act 2004. http://www.legislation.gov.uk/ukpga/2004/30/pdfs/ukpga_20040030_en.pdf. Accessed 03 July 2015.

UK Parliament. 2004b. Human Tissue Act 2004 Explanatory Note. http://www.legislation.gov.uk/ukpga/2004/30/notes. Last consulted 20.06. Accessed 03 July 2015.

Willimas, George. 1992. *Natural selection: Domains, levels and challenges*. New York: Oxford University Press.

Zins, Chaim. 2007. Conceptual approaches for defining data, information and knowledge. *Journal of the Association for Information Science and Technology* 58(4): 479–493.

What's Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics

Bryce Goodman

Abstract Advances in full genome sequencing have led to a practical and ethical conflict between protecting genetic privacy and large-scale genomic research. This chapter concerns the value of genetic privacy, and consists of both a negative and positive claim. The negative claim is that genetic privacy is not intrinsically valuable, and that the barriers to genomic research posed by an unqualified right to genetic privacy are not justified. The positive claim is that genetic research is supported by the principle of respect for autonomy.

1 Introduction

Genetic research is at a cross-roads. Less expensive genetic sequencing and the application of computational methods to resulting datasets have led to new fields of genomic research.[1] This confluence of "big data" and genetics has created an unprecedented demand for access to large sets of genetic information. The demand is being met by the creation and expansion of population bio-banks, which store genetic and non-genetic information (e.g. medical history, demographics, etc.). The information stored in bio-banks can be used in multiple *genome wide association studies* (GWAS), which analyze statistical patterns of genetic variation to find genetic correlates for common diseases. These developments suggest that, in the not so distant future, genomic research will have profound applications in both clinical medicine and public health (Visscher et al. 2012).

[1]The sum total of information contained within an organism's DNA is its genome; a person's genome is of the human genotype, but "genotype" can also refer to classes of genomes *within* the human genotype. The terms genomic and genetic research refer to research conducted using most or all of the information contained within the subject's genome.

B. Goodman (✉)
Oxford Internet Institute, St. Cross College, University of Oxford, 61 St Giles',
Oxford OX1 3LZ, UK
e-mail: bwgoodman@gmail.com

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_7

However, some see the demands of large-scale genomic research, which depends upon ready access to vast amounts of genetic information,[2] as a threat to genetic privacy. Genetic exceptionalism[3] is the view that genetic information is "sufficiently different from other kinds of health-related information that it deserves special protections or other exceptional measures" (Suter 2001, 669). Proponents of genetic exceptionalism demand more stringent rules for accessing genetic information as compared to medical information generally. This demand is grounded in the claim that genetic privacy is intrinsically valuable because genetic information is analogous to the contents of a person's "probabilistic future diary" (Anderlik and Rothstein 2001; Annas et al. 1995, i). Proponents of this view recognize that stricter regulations may hinder, or indeed obstruct, the progress of large scale genetic investigations. However, according to this view, these impediments are ethically justified because the right to genetic privacy trumps the value of genomic research.[4]

This chapter claims that proponents of genetic exceptionalism are mistaken on two counts: first, in positing a necessary connection between genetic privacy and respect for autonomy, and second, in embracing a reductionist and determinist conception of genetic causation. Contrary to both of these views, genetic privacy is not intrinsically valuable, and may actually conflict with the obligation to protect and promote autonomy. Consequently, there is a need to seriously re-think the value of genetic privacy, both in terms of measures for its protection and its role in discussions of how to regulate access to genetic information.

Section 2 introduces genetic exceptionalism and the conflict between the right to genetic privacy and genomic research. Section 3 considers challenges to the right to privacy, and shows how these objections motivate proponents of genetic privacy to claim that it is intrinsically valuable as a facet of autonomy. Section 4 queries and rejects the putative "intrinsic connection" between privacy and autonomy, and argues that an additional reason is needed to prove that genetic privacy is connected to autonomy. Section 5 evaluates and rejects the connection between genetic information and autonomy asserted by genetic exceptionalism. Section 6 combines the conception of autonomy presented at the conclusion of Section 4 with the view of genetic information developed in Section 5 to argue that an unqualified right to genetic privacy is both unsupported by, and may conflict with, the principle of respect for autonomy. Proponents of genetic exceptionalism focus exclusively on the potential harm that can arise from unauthorized release of genetic information and so neglect the fact that genetic research provides individuals with information that enhances their autonomy.

[2] Information derived through genetic sequencing.

[3] Although not all proponents of genetic privacy claim to be proponents of genetic exceptionalism, the two positions tend to go together, and are mutually supportive. Further, it seems plausible to assume that anyone who is a proponent of genetic privacy believes that genetic information is exceptional in *some* way.

[4] The concept of rights as trumps is from Dworkin (1978).

2 Genetic Exceptionalism and the Right to Genetic Privacy

This section introduces the thesis of genetic exceptionalism, which maintains (1) that genetic information is ethically distinct from other types of medical information and (2) that it requires special protections. In the context of genetic exceptionalism, the right to genetic privacy is viewed as a fundamental and basic individual protection which trumps the interests of genetic research and imposes a blanket obligation not to access an individual's genetic information without first obtaining both explicit and informed consent (Caplan 2009; Roche et al. 1996). Both requirements pose specific challenges to bio-bank and data-sharing projects, which are necessary for the continuation of large-scale genomic research.

2.1 The Origins of Genetic Exceptionalism

Arguments in favor of a "special legal status" for genetic information—information derived through genetic sequencing – in the United States can be traced back to a 1984 report from the President's Commission on Bioethics, which declared that genetic information should not be released without first obtaining "explicit and informed consent" (Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research 1984, 303).

During the early 90s, advances in genetic research and the discovery of genetic correlates for common diseases motivated the introduction at both state and Federal level of numerous bills aimed at banning potential forms of genetic discrimination, e.g. denial of health coverage or employment on the basis of a person's genetic traits (Parthasarathy 2004, 243). Whereas genetic tests had once been limited both in terms of scope and accuracy, the flurry of research development in the 90s began to put pressure on mainstream practice to respond to the clinical promises of genetic research (Bove et al. 1997; Burgess et al. 1997; Dickens et al. 1996).

From the beginning, advocacy groups worked to promulgate an understanding of genetic information as both ethically distinct from other types of medical information and intrinsically private (Parthasarathy 2004, 243). For example, after the discovery of the BRCA 1 gene mutation and its link to certain types of genetically caused breast cancer, the NIH funded the Hereditary Susceptibility Working Group of the National Action Plan on Breast Cancer (NAPBC). The Group released a report stating, "genetic information is distinct from other types of medical information because it provides information about an individual's predisposition to future disease"[5] (Hudson et al. 1995, 392).

[5]It is worth noting that this statement is, strictly speaking, false. For example, a person's body mass index may be a much better predictor of diabetes than any information derived from genetic code.

The statement was met with public alarm; by March 1995, 118 bills had been passed in the United States at both the Federal and state level in response to growing fears over the possible misuse of genetic information, particularly in the area of health insurance (Mulholland 1998).

2.2 Genetic Privacy Verses Genetic Discrimination

Genetic exceptionalism sees genetic information as intrinsically valuable, that is, something to be protected irrespective of consequences. An alternative is to argue that genetic privacy ought to be protected as a means of minimizing genetic discrimination, the practice of withholding services (e.g. health care) or otherwise treating individuals less favorably on the basis of information derived from their DNA.[6] In this case genetic privacy has instrumental or contingent value – it is valuable insofar as it lessens or prevents genetic discrimination.

Concerns relating to genetic discrimination have, historically, been a significant source of motivation for stricter genetic privacy laws (Everett 2004; Diver and Cohen 2000; Collins and Watson 2003). For example, a 2004 United Nations Declaration addressing both genetic privacy and genetic discrimination asserts a common rationale for the two rights, "*Recognizing that* revealing genetic information belonging to individuals without their consent may cause harm and discrimination against them . . . " (United Nations Economic and Social Council 2004).

However, genetic discrimination is easily distinguished from genetic privacy. Any unauthorized access to genetic information violates genetic privacy, regardless of whether genetic discrimination takes place. Further, a number of studies suggest that there is no empirical connection between laws protecting genetic privacy and cases of genetic discrimination (Hall and Rich 2000; Nowlan 2002), and some authors have disputed the claim that genetic discrimination is an actual practice *tout court* (Taylor et al. 2003; Green et al. 2015).

Furthermore, even a policy that specifically addresses concerns over *genetic* discrimination need not limit *research* access to genetic information. For example, while the United States' 2008 *Genetic Information Non-Discrimination Act* (GINA) protects individuals from genetic discrimination, it does not stipulate any prohibitions or conditions for the use of genetic information in biomedical research.[7]

From a legal perspective, a prohibition on genetic discrimination is entailed by a general prohibition against discrimination on the basis of *any* medical information; such policies need not presuppose any of the claims of genetic exceptionalism. In their recent review of the impact of GINA on genetic discrimination, Green et al. (2015) report that, in 2013, of the 333 claims related to genetic discrimination in

[6]This definition is offered as an attempt to distinguish genetic discrimination from discrimination on the basis of traits (e.g. race) that have a genetic component.

[7]For international laws against genetic discrimination, see Motoc (2009, 222–246).

employment (compared with more than 90,000 in other areas), most also included claims relating to non-genetic discrimination.

Finally, as noted earlier, the argument that connects genetic privacy with genetic discrimination sees the former as having derivative or instrumental value. In this case, the right at stake is freedom from genetic discrimination, and not genetic privacy. If circumstances were to arise where the connection between privacy and discrimination no longer obtained, a proponent of this view would be forced to offer some additional reason for why genetic privacy ought to be protected. For proponents of genetic exceptionalism, however, merely accessing an individual's genetic information without obtaining consent is wrong, regardless of consequences.

2.3 The Right to Genetic Privacy

The original and perhaps most influential articulation of the right to genetic privacy is the *Genetic Privacy Act: A Proposal for National Legislation* (*GPA*), which was drafted in 1995 by Annas and colleagues at the University of Boston. Its authors write:

> The overarching premise of the Act is that no stranger should have or control identifiable DNA samples or genetic information about an individual unless that individual specifically authorizes the collection of DNA samples for the purpose of genetic analysis, authorizes the creation of that private information, and has access to and control over the dissemination of that information (Roche et al. 1996, Introduction).

The *Act* is premised on the notion that genetic privacy is intrinsically valuable (Everett 2004). According to this view, individuals have a *special right* to genetic privacy.[8]

The *Genetic Privacy Act* prohibits both data-sharing and derivative use of samples without first obtaining explicit (e.g. verbal or written) and informed (e.g. project-specific) consent. The concept of informed consent as an ethical requirement for medical research was first developed during the Nuremburg trials and further developed in the Declaration of Helsinki (Manson and O'Neill 2007). At Nuremburg, informed consent was used to distinguish between ethically acceptable research and the experiments conducted by Nazi's during World War II (Annas et al. 1992). Since many forms of medical research pose a risk of harm to participants, this factor alone was insufficient to separate the actions of the Nazis from the actions of legitimate researchers.

The prosecution argued that what made the Nazi case distinctive is that the participants were coerced; they did not give their free, explicit and informed

[8]Such rights are "special" insofar as they impose additional restrictions not already entailed by regulations for medical information generally. See, for example, Annas (1993, 1995, 1999), Annas et al. (1995), Rothstein (2005).

consent to participate in the experiments they were subjected to. In the course of this argument the two specific requirements of informed consent came to be developed. First, the consent has to be explicit—an absence of objection is insufficient grounds for assuming that a participant consents. Thus the onus is on the researcher to show that consent was obtained, "preferably in writing" and otherwise "formally documented and witnessed" (World Medical Association 2001, Sec. 22). Second, before consent can be obtained, "each potential subject must be adequately informed of the aims methods, course of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail" (World Medical Association 2001, Sec. 22).

In the context of large-scale genomic research, informed consent poses a number of practical and conceptual challenges. The first arises from the obligation to obtain consent that is explicit. The purpose of bio-banks is to provide easily accessible genetic information to qualified research groups and thereby reduce the high cost of collecting data for large scale studies. Trying to obtain explicit written or witnessed verbal consent from bio-bank participants, which could number in the hundreds of thousands, would obviate the entire point of such endeavors (Ursin 2008, 267–269; Hansson 2009, 9; Lunshof et al. 2008).[9] The obligation to obtain *informed* consent deepens this dilemma, since the information gathered by bio-banks is intended for use in a range of future projects that are not specifiable at the time when individuals consent to participation.[10]

Proponents of the rights outlined in the *GPA* acknowledge that such legislation may hinder the progress of genetic research. At the same time, they maintain that these restrictions are ethically required because researchers working with genetic information "are handling the most intimate and personal information that will ever exist" (Goerl, Hyer, and Farkas 1997, 86). Caplan (2009) argues that the cost of protecting genetic privacy cannot be weighed against the potential benefits of genomic research:

> Yes, the more of us contribute our DNA and correlated personal information to the pursuit of genomic science the better it might be for science and medicine . . . [but] the correct course in the emerging age of genomics is not to dispose of privacy . . . Rather, we must seek ways to strengthen and maximize the opportunity for you to invoke privacy while finding ways to conduct genomic research that can protect it.

[9] Specific challenges are also presented by research involving subjects who are typically viewed as not able to consent, e.g. children (Gurwitz et al. 2009) and archival samples (Steinberg et al. 1995; Beskow et al. 2001; Bathe and McGuire 2009).

[10] There is currently a debate over whether policies of "broad consent," where participants would agree to future unspecified use of their genetic information, is an ethically acceptable alternative. Critics argue that "There is no such thing as 'general informed consent'" and that "it is misleading to use the notion of informed consent for participation in research that is unforeseen and has not been specified in a research protocol." (Arnason 2004, 41; Caulfield 2007; Greely 2007). For defense and further discussion, see Hansson et al. (2006), Beskow et al. (2001), Lunshof et al. (2008).

The right to genetic privacy is, according to this view, a *moral constraint* upon genomic research.[11]

2.4 The Need to Re-think Genetic Privacy

Although it remains highly influential, the *GPA* was never formally adopted. The majority of policies have taken a less rigid approach to regulating genetic information (Eriksson and Helgesson 2005). Some authors believe there ought to be a balance struck between research and privacy interests, and encourage strict enforcement of privacy guidelines, and new methods for de-identification and data-anonymization.[12] According to this view, if the benefits of allowing access to genetic information for the purposes of research are sufficiently high, they may outweigh some of the "harms" that would be caused by the violation of genetic privacy (Gurwitz et al. 2009, 819).

While this position may, at first blush, seem to strike a balance between privacy and research, it is both practically limited and does not address fundamental questions regarding the nature or value of genetic privacy.

First, relying on stricter data-coding policies alone may be for naught; recent studies have shown that ensuring complete anonymization of genetic data is practically impossible in genomic research (Gymrek et al. 2013).[13] This raises the question of whether large scale genetic research is permissible even if genetic privacy cannot be absolutely guaranteed, a question that can only be addressed following a careful consideration of the value of genetic privacy.

Furthermore, there may be a practical conflict between the level of anonymization and the quality of research outcomes: depending on what information is removed, attempts to anonymize genetic data may diminish the scientific utility of the collected data. For example, ethnographic and geographic information about a cohort may be simultaneously useful for researchers whilst increasing the feasibility of re-identification.

[11]The idea of rights as moral constraints is from Nozick (1974, sec. Moral Constraints and Moral Goals).

[12]Information is considered "de-identified" once any uniquely identifying data is removed, e.g. name, driving license, national insurance number, etc. Information that is de-identified may or may not be re-identifiable; coding allows for de-identification during research or data-sharing, but may also allow for the re-identification of genetic information and DNA samples. Such information is considered de-identified but remains potentially identifying; information that is no longer potentially identifying (e.g. where all linkages to the individual have been permanently destroyed) is considered anonymous. See McGuire et al. (2008).

[13]Controversy over the plausibility of maintaining anonymity first arose after the publication of Homer (2008). The report showed that information previously considered fully anonymous was, in fact, potentially identifying. The report's findings have been contested, but prompted both the Wellcome Trust and NIH to withdraw public access to certain anonymized genetic data-sets. See also Gymrek et al. (2013).

There is also a tension between anonymization and ensuring individuals retain control over their genetic information (Rothstein 2010). One of the requirements of informed consent is that subjects have the right to withdraw from participation at any point during the study (World Medical Association 2001, Sec. 22). However, as Bathe and McGuire (2009, 713) note, "anonymization precludes any influence the donors have on the use of their samples," and makes it practically impossible to re-contact participants for obtaining consent or returning clinically significant findings.

Finally, because data-sharing between research groups is often required by funding bodies, effectively regulating access will depend upon the existence (and acceptance) of uniform international guidelines (Kaye et al. 2009a, 332). Harmonized regulations are currently lacking and, as one survey notes, where guidelines *do* exist, they "reflect a debate that is characterized by perplexity and controversy" (Mauron and Boggio 2005, 3).

The "perplexity and controversy" that characterizes existing frameworks is not for lack of effort, but reflects the contentious debates over which *moral principles* ought to guide the practices of data-collection, storage and use (Elger and Caplan 2006, 664). As will be discussed in the following section, the lack of clarity in genetic privacy regulations may stem from fundamental challenges with the nature and value of privacy itself.

3 From Privacy to Autonomy

This section introduces practical and conceptual challenges for defining the scope of the right to privacy. When privacy is left overly broad, it is not clearly distinguishable from other values or rights; on the other hand, when privacy is given an overly narrow scope, its value ceases to be significant. Similar challenges face the right to genetic privacy, and motivate the argument that connects genetic privacy with the value of autonomy.

3.1 The Nature and Value of Privacy

According to the Oxford English Dictionary, something is *private* if it relates to or affects "a person or a small intimate body or group of persons apart from the general community; [it is] individual, personal" (Oxford English Dictionary 2010). The distinction between a private and public sphere goes back at least as far as Aristotle and the division between *polis* and *oikos*—the former denotes a place where citizens engage in public undertakings, whereas the latter is the area occupied by domestic life (Ursin 2008, 272). However, the Greek distinction does not explain the view that privacy is intrinsically *valuable* and of particular importance for *individuals*:

private life, while distinct from public engagement, is not of ultimate importance in Aristotle's conception of the human good.[14]

An explicit treatment of the value of privacy is not only absent from Aristotle's description of the virtuous life; privacy also receives little mention from the major thinkers in the liberal tradition.[15] The origin of privacy as a moral concept[16] is more readily located within Judeo-Christian theology. Konvitz (1966, 276) notes the prominent role that privacy and concealment have played in the development of Western normative and theological concepts: "Mythically, we have been taught that our very knowledge of good and evil—our moral nature, our nature as men—is somehow, by divine ordinance, linked with a sense and a realm of privacy." Konvitz is alluding to the Biblical tale of Adam and Eve, wherein shame—and moral knowledge—coincide with the need for privacy. The story suggests that the concern for privacy is a fundamental part of what distinguishes civilized interaction from interaction amongst animals and barbarians; from this perspective, the need for privacy is intimately bound with an individual's status as a dignified moral agent (Etzioni 2000, 166–167; Ursin 2008, 269).

3.2 Origins of the Right to Privacy

The idea that privacy requires specific legal protections, however, has its roots in American legal history. It is here that one finds one of the first instances of privacy as a value that imposes restrictions on *others*, along with an attempt to distinguish privacy from related, but distinct, concepts, such as liberty and property ownership.

The legal right was first introduced in an article by the United States Justices Warren and Brandeis, which responded to privacy challenges posed by the advent of photography and mass publication:

> Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the new demands of society. Thus, in very early times, the law gave a remedy only for physical interference with life and property . . . now the right to life has come to mean the right to enjoy life–the right to be let alone (Warren and Brandeis 1890, sec. Introduction).

[14]Nevett cautions against conflating the Greek *oikos* with the modern notion of a "private sphere." (Nevett 2001, 5). This view has been contested by those who believe that Aristotle *did* place distinctive value on the privacy afforded by family life, e.g. (Salkever 1977). However, even if Aristotle did believe the *oikos* represented a special sphere for family life, there is no evidence to suggest he believed *individual* privacy (e.g. seclusion from public life) was a value, and certainly not a value that could compete with or trump the good of the *polis*. Cf. Aristotle's discussion of political science at Book I Chapter 2 of the *Nicomachean Ethics* (Irwin 1985, lines 1094b 9–10).

[15]McCloskey (1980, 17–18) calls attention to this often-ignored fact and notes that neither Locke, Rousseau, Kant nor Mill discuss privacy as a moral concept.

[16]E.g. Privacy as a source of moral obligations.

The Justices sometimes write of the right to privacy as "the right to one's personality." Accordingly, the Justices argue that "the protection afforded" by a right to privacy "is not confined . . . [to] any particular medium or form of expression has been adopted, nor to products of the intellect" (Warren and Brandeis 1890). The right to privacy thus establishes a broad *prima facie* obligation[17] not to intrude upon an individual's "personal sphere."

3.3 Challenges for the Right to Privacy

Two related conceptual concerns challenge the view that, generally, privacy should be regarded as a basic right. The first pertains to the scope of privacy: if privacy lacks definite scope, the obligation to respect a person's privacy is not meaningful. Responding to this challenge, some authors invoke territorial and spatial metaphors. Drawing upon a connection with territorial sovereignty, Feinberg argues that privacy denotes "a certain amount of 'breathing space' around one's body" (Childress and Beauchamp 2001, 297).

While the metaphor of a "personal sphere" or "intimate zone" may be emotionally evocative, it still fails to provide a concrete principle that could determine, in practice, whether and when a person's *privacy* is genuinely at stake. Thomson (1975, 295) argues that if the right to privacy is understood as the right to be left alone, its scope is obscure and encompasses far more than privacy intrusions. One way we might fail to let someone alone is by hitting him, however "doing [so] should surely not turn out to violate his right to privacy. Else, where is this to end? Is *every* violation of a right a violation of the right to privacy?"

A second challenge emerges from the claim that the value of privacy is basic and universal. In practice, different societies have radically divergent or even conflicting views about what ought to be regarded as private (Whitman 2004). This diversity of perspectives is not only anthropologically interesting, but ethically significant – it suggests that what counts as private is the product of both individual preferences and cultural norms, grounded in specific traditions. Whitman (2004, 1152) observes:

> If privacy is a universal need that gives rise to a fundamental human right, why does it take such disconcertingly diverse forms? This is a hard problem for privacy advocates who want to talk about the values of "personhood."

Those seeking to justify the need for privacy as a fundamental human right, genetic or otherwise, face a dilemma with two horns: if privacy is left overly broad it ceases to have any distinguishing content, and if privacy is given an overly narrow specification, it is not a universal value.

[17] A *prima facie* obligation is to be distinguished from an *actual* or *ultimate* obligation. An agent may be exempt from a *prima facie* obligation depending upon other competing obligations. For example, the Justices note that the obligation to respect an individual's privacy may be overruled or suspended when there is sufficient "public interest" in that individual's affairs. See Ross and Stratton-Lake (2002).

3.4 Challenges for Genetic Privacy

These challenges are amplified in the case of genetic privacy. "Genetic information" is, in its broadest sense, any information contained within an organism's genome. Such information may or may not be informative or clinically significant: a sequence of "un-interpreted" nucleotides is not informative because the raw data obtained through genetic sequencing must be *interpreted* before it is intelligible.[18] However, many genetic privacy laws, including the *GPA*, do not recognize a distinction between information that is genetic (e.g. derived from a person's DNA) and information that is both genetic and genuinely informative (e.g. interpreted) (Sankar 2003; Manson and O'Neill 2007). If "genetic information" simply refers to information about a person that is, in some sense, linked to that person's genes, it is no longer a distinctive category: skin, eye and hair color all have genetic determinants, but one need not access a person's DNA to say what color eyes he or she has.[19]

The problem for proponents of genetic exceptionalism is that, oftentimes, genetic information is not exceptional. In an article on regulating access to bio-banks, Annas (1993, 2348) argues that stricter policies are needed because "... genetic information has been grossly misused in the past, especially in the eugenics movement and Nazi Germany's program of racial hygiene." However, within the context of the article, Annas (1993, 2346) is addressing "information derived from the DNA sample." Annas' argument rests upon an equivocation between two very distinct senses of genetic information. The Nazis did not have access to information *derived from* genetic analysis: consequently, they would not have had access to the information protected by a right to genetic privacy, nor would such a right, in principle, limit access to the information the Nazis *did* have. The error is in conflating information that says something *about* someone's DNA with information that we have derived *from* someone's DNA.

Another challenge arises when "genetic privacy" is used to denote issues bearing little or no logical connection. For example, Allen attempts to carve out four distinct types of medical privacy concerns and claims that all are at stake in genetic privacy (Rothstein 1999, Chap. Genetic Privacy):

1. Informational privacy concerns access to personal information
2. Decisional privacy concerns governmental and other third party interference with personal choices
3. Physical privacy concerns access to persons and personal spaces
4. Proprietary privacy concerns the appropriation and ownership of interests in human personality

[18] A genetic variation is "interpreted" when correlated with a clinically significant trait.

[19] Furthermore, some additional reason would be needed to explain the purported ethical significance between coming to know a person has blue eyes because they have shared their DNA versus seeing them on the street.

While the taxonomy presented by Allen may be illuminating for discussions of medical privacy generally, decisional, physical and proprietary privacy concerns are, at most, contingently related to genetic information.

In medical ethics, decisional privacy appeals to "privacy as an [American] constitutional/legal concept, especially the post-1965 constitutional right of privacy [which] emphasizes decisional privacy, that is, the freedom to decide and to act in public or private as one deems appropriate, without government interference" (Margulis 2003, 244).[20] Provided individuals are not coerced by governments into participating in genetic bio-banks, it is difficult to see how decisional privacy is of any relevance whatsoever. Furthermore, it is debatable whether the concept of decisional privacy, in the sense provided, can be meaningfully distinguished from either liberty or autonomy (Ursin 2008, 270); while this would not, of itself, show that concerns grouped under the heading of decisional privacy are unimportant, it calls into question the distinctiveness of privacy rights.

Concerns over physical privacy are out of place in the context of bio-banks. Large scale genetic research is conducted using digitally stored genetic information, so there is no physical contact between researcher and subject. Furthermore, DNA analysis may be conducted using recently collected or "archival" bio-specimen samples, and thus even the sequencing need not necessarily involve any further physical contact with the participant (Bathe and McGuire 2009). Even when contact *is* necessary for collecting genetic material, obtaining suitable biological specimens (e.g. saliva, hair, etc.) can be done remotely and is significantly less physically intrusive than many other forms of medical testing.

The notion of proprietary privacy is often connected with the theory of ownership developed by John Locke. According to this view, Locke "provided a philosophical justification of the right to property which extended to matters of privacy by promoting the idea that we do not merely exist in our bodies, but also *own* them" (Ursin 2008, 276).[21] However protections afforded by a right to privacy differ in kind from those entailed by property rights: in the former, the object protected is distinguishable from the right-bearer, whereas in the latter the right bearer *is* the subject of protections (Suter 2003). The property analogy thus implies that the thing intruded upon is distinct from the owner. This distinction is untenable in the case of "genetic information," because an individual cannot, in either a logical or legal sense, ever "give up" his or her genetic code (Rao 2007).

[20] In the United States, this notion played an important role in the argument that a right to privacy prohibited state bans on abortion rights.

[21] However, it is important to note that Locke's conception of self-ownership is not unqualified; according to Locke, individuals are *stewards* over their bodies, and have obligations stating how they may or may not treat their "property." As Ryan notes, "the starting point of an adequate interpretation of Locke's [concept of self-ownership] begins with the general principle that human beings have only those rights in themselves or over anything else that are required to enable them to achieve the purposes for which God established the world and created them." (Ryan 1994, 243)

Another reason not to countenance a property-based approach to genetic privacy is that it runs the risk of drawing attention away from informational privacy and towards the increasingly marginal issue of physical privacy.[22] In other words, modeling genetic privacy solely on the basis of property ownership may fail to capture the full range of claims associated with a right to genetic privacy, which pertain to *both* physical and non-physical access (e.g. access to genetic information).[23] Finally, this theory fails to explain how information can be "owned" by an individual. Frey writes:

> Suppose that I have a certain illness and do not want the fact of my having it to be released: how do I acquire a private property right in this information in such a way that the fact that I have this illness becomes, as it were, mine? For notice, nothing less than owning the fact—as it were, owning the information that that [*sic*] I have the illness- would be compatible with others learning this information but violating no private property of mine... (Ursin 2008, 277)

The notion that information is individually owned is particularly problematic in the case of genetics, where information that is derived from one person's DNA may have significant implications for family members. For example, a child's decision to get tested for the Huntington gene, which is inherited, clearly provides information that is relevant to that child's parents. If the child tests positive, this result will indicate that a parent carries the gene as well; given such a case, asserting that either the child or the parent "owns" the information is patently absurd. As Sommerville and English (1999, 144–145) note, in the case of genetics, "few decisions are entirely personal."

Thus enforcing strict individual genetic privacy rights may prove impossible because not all of the information obtained through DNA analysis is uniquely related to just one person. Further, as the example shows, one individual's decision to allow or deny access to her genetic information may greatly affect the "genetic privacy" of another.[24]

3.5 Retreat from Privacy and the Argument from Autonomy

Thomson (1975) argues that the right to privacy should not be recognized as protecting a single, unified interest, but rather a cluster of interests that may or may not be logically related.[25] A person may have rights not to be interfered with, according to Thomson, but not because of a right to privacy. Rather, Thomson claims

[22] For further discussion of the importance of, and transition towards, digital bio-banks, see Church et al. (2009), Kaye et al. (2009a, b), Mauron and Boggio (2005), Krawczak et al. (1999).

[23] The limitations of a property based approach to privacy rights motivated the original articulation of the right to privacy. See Warren and Brandeis (1890).

[24] For further discussion see Doukas and Berg (2001).

[25] In contrast to, for example, Scanlon (1975, 316).

that the right to privacy can always be analytically reduced to other, more basic rights (Thomson 1975, 312). Thomson concludes that the right to privacy has no independent moral grounding, and so does not merit recognition as a distinctive source of legal or moral obligations.

Other authors argue that the value of privacy *for an individual* is always dependent upon other values that may require or be promoted by privacy. For example, McCloskey (1980, 38) writes:

> To be plausible, any account of privacy must explain privacy as something distinct . . . and as something *sui generis*. When so explained, privacy of itself does not provide grounds for believing men as men to possess a basic moral right to privacy. Any right to privacy must be based on other rights and goods.

Although neither Thomson nor McCluskey's position is universally accepted,[26] Allen notes that while "a few theorists depict privacy as denoting a basic human good whose value is intrinsic," most proponents of genetic privacy argue that the right to genetic privacy appeals to "ideals of personhood consisting of rational, self determining, morally autonomous individuals" (Rothstein 1999, 35). Gostin (1995) also asserts that respect for autonomy provides the normative grounding for privacy. This approach is adopted by Anderlick and Rothstein (2001, sec. Arguments in favor of protecting privacy):

> Because genetic information is connected to personal and group identity, protecting privacy of genetic information is an important individual and social priority. Genetic privacy has intrinsic value as a facet of autonomy, and respect for autonomy implies a duty to respect the genetic privacy of others. Within a legal framework, genetic privacy must be considered a fundamental right.

The next two sections question the putative connection between autonomy and privacy presupposed by this argument for genetic privacy.

4 Autonomy and the Right to Privacy

This section concerns the alleged link between the concepts of privacy and autonomy. Some authors maintain that autonomy consists in the absence of external influence and that privacy is a necessary condition for the development and preservation of an "autonomous self." However, this section will argue that privacy and autonomy are only contingently related: autonomy depends upon conditions both external and internal to an agent, whereas privacy primarily concerns the absence of certain external factors. Consequently, the argument for genetic privacy requires some additional reason or argument to link genetic privacy to autonomy.

[26] For a response to Thomson, see Scanlon (1975). For a more general reply, see Parent (1983b).

4.1 The Nature and Value of Autonomy

Autonomy comes from the Greek *autos* (self) and *nomos* (rule of law). A state is autonomous when its government determines its actions: in Ancient Greece, a city had *autonomia* when it was not subject to the rule of another state (Dworkin 1988, 108). In biomedical ethics, the analogy with self-governance suggests that autonomy depends upon making and following self-issued judgments:

> The autonomous individual acts freely in accordance with a self-chosen plan, analogous to the way an independent government manages its territories and sets its policies (Childress and Beauchamp 2001, 58).

When authors appeal to personal autonomy as a source of obligations they do so on the grounds that the ability to exercise self-determination, e.g. through freedom of expression, is *valuable* for individuals, and that one's autonomy can be impeded by external factors, e.g. incomplete or mistaken understanding of the conditions in which one operates, or the interfering actions of others.

In *On Liberty*, JS Mill argues powerfully for a connection between choosing for oneself and the cultivation of an individual's character; the ability to act autonomously is intimately bound with our status as intelligent, and not merely imitative, beings. For Mill, the value of self-determination consists not only in the fulfillment of certain desires, but also in the cultivation of authenticity (Mill 1869).

Mill's writings on autonomy encourage us to distinguish between autonomy as a human *capacity*, and as a feature of *actions*: an autonomous agent is one who is capable of acting autonomously, but may not always do so. Similarly, Beauchamp and Childress (2001, 63) claim that personal autonomy concerns the capacity of a person to make and follow self-given plans and is, "at a minimum, self-rule that is free from both controlling interferences by others and from limitations, such as inadequate understanding, that prevent a meaningful choice." Inadequate understanding can hinder autonomy because autonomous choice is not just choice in the absence of external influence.

4.2 The Argument from Autonomy

The comparison between an autonomous individual and independent state has led some theorists to model personal autonomy upon the political ideal of sovereignty. A state which is sovereign has "supreme dominion, authority, or rule" over its territories. According to proponents of "individual autonomy," the autonomous agent is, first and foremost, *detached* from influences that may interfere with his capacity to exercise choice.[27] Thus the individualistic conception of autonomy

[27] Cf. "To regard himself as autonomous in the sense I have in mind, a person must see himself as sovereign..." (Scanlon 1972, 215) "The autonomous man, insofar as he is autonomous, is not

equates the state of being autonomous with *negative liberty*[28] or substantial independence from external influence. When a premium is placed upon independence from outside control, the principle of respect for autonomy becomes primarily a positive requirement to obtain informed consent and, negatively, a restraint on what others can do to an autonomous person.

According to "control-based" accounts, "privacy" refers to an individual's ability to *control* and *authorize* certain types of access. Westin (2003, 431), for example, defines privacy as "the claim of an individual to determine what information about himself or herself should be known to others," and Margulis (2003, 245) writes that privacy is "control over transactions between person(s) and other(s), the ultimate aim of which is to enhance autonomy and/or minimize vulnerability, [which] usually entails limits on or regulation of access to self."

Taken together, control-based accounts of privacy and conceptions of individual autonomy provide a firm grounding for the connection between respect for autonomy and privacy. For example, Nagel (1998, 12) claims that the individualist conception of autonomy is modeled after the "liberal idea [that] in society and culture as in politics . . . no more should be subjected to the demands of public response than is necessary for the requirements of collective life." Nagel (1998, 7) also defends a connection between privacy and autonomy; he argues that the individual can only resist the influences of society, and maintain an "inner" or "true" self, through concealment and selective disclosure. According to Nagel, privacy is a necessary condition for autonomy because the very activity of social interaction is a potential encroachment upon the domain over which one's "true" self is sovereign. From this vantage, privacy as the ability to control access to one's "autonomous self" appears *intrinsic* to autonomy (Kupfer 1987).

4.3 Objections to the Argument from Autonomy

However, upon closer examination, this connection between autonomy and privacy appears frayed. It is implausible that respect for privacy is derived from respect for autonomy because we can respect the privacy of a person who lacks autonomy. Parent (1983a, 326) offers the case of a comatose patient. The patient clearly lacks autonomy. But while there is no sense in which one could respect the patient's

subject to the will of another." (Wolff 1998, 14) "I am autonomous if I rule me, and no one else rules I." (Feinberg 1973, 161). The term "individual autonomy" is from O'Neill (2002).

[28]The term is from Berlin's "Two Concepts of Liberty," in which he writes that "liberty in this sense is simply the area within which a man can act unobstructed by others.'" Berlin claims that interference with freedom is limited to cases of coercion, which includes an intentional aspect and is distinct from cases where a person is limited by what she can or cannot do by either a lack of capacity or natural laws. The liberty associated with the individualistic conception of autonomy is arguably more expansive; as we will see, interference with individual autonomy includes even cases where others may not *intend* to diminish a person's autonomy.

ability "to determine what information ... should be known by others ... efforts to safeguard his privacy still make perfectly good sense, and can be entirely successful." If privacy is understood in terms of control over information, however, there would seem to be no way of accounting for such an obligation, or to explain how one might protect the privacy of individuals who lack or have a diminished capacity for self-control. Similarly, there are cases where an individual's privacy may be compromised without impacting autonomy, e.g. forgetting to shut one's blinds (Taylor 2002) or speaking too loudly on one's cell phone in a crowded bus. Weinreb (2009, 41–42) dismisses the putative claim that privacy is a necessary condition for autonomy because "in each instance [privacy] *directly* concerns what *others* know. Although the context in which the acts has changed, the person whose privacy is invaded is, *qua* actor, the same after the invasion as before; or if he is not, it is because of *his own* reaction to the fact that others have acquired the information in question about him." Similarly, Dworkin (1988, 21) claims that, in general, privacy may be diminished without affecting autonomy.

In the context of genomic research, there is no obvious reason why allowing researchers to access an individual's genetic information should necessarily change that person's decisions, actions or values. It is infinitely more likely that there is no contact between a participant and researchers, and that participation in research does not affect a person's sense of being autonomous in any way. To be clear, genetic discrimination could have a direct impact on an individual's autonomy. However, as discussed in Sect. 2.2, there is a clear distinction between a right to not be discriminated against and the right to genetic privacy espoused by supporters of genetic exceptionalism: privacy concerns access to genetic information, whereas discrimination concerns actions that may or may not be based upon, or take place following, access to genetic information. The argument for genetic exceptionalism requires an additional reason to show that genetic privacy is *intrinsically* valuable as a facet of autonomy.

5 Genetic Information and Autonomy

Some authors argue that genetic privacy is related to autonomy because of a unique connection between individuals and the information derived from their DNA. This view is supported by the theses of genetic reductionism and genetic determinism, both of which have played a prominent role in shaping the discourse of genetic privacy through evocative metaphors. However, this section will argue that both reductionism and determinism are empirically false, and that the metaphors they offer fundamentally misconstrue the nature of genetic causation and the purposes of genetic research.

5.1 Metaphors of Genetic Exceptionalism

The authors of the *Genetic Privacy Act* describe genetic information as analogous to the contents of an individual's "future diary" (Annas et al. 1995, ii):

> A diary is perhaps the most personal and private document a person can create. It contains a person's innermost thoughts and perceptions and is usually hidden and locked to secure its secrecy. Diaries describe the past. The information in one's genetic code can be thought of as a coded probabilistic future diary because it describes an important part of a unique and personal future.

The metaphor is effective because the information contained in a personal diary may be important to protect irrespective of whether others intend to harm the author in any way. This intuition might be supported by two arguments. First, a personal diary is a piece of private property, and one may acknowledge that, along with rights of ownership, come rights to decide who may or may not have access; following Nozick (1974, 171–172), one might think that these rights exist irrespective of whether they protect some interest that a person has in keeping his or her property private from others. One might argue that, if genetic information is analogous to the contents of a personal diary, it ought to be protected irrespective of whether an individual will be less able to act autonomously as a result.[29]

A second argument supported by the future diary analogy carries the connection between genetic information and autonomy further. If genetic information is analogous to the contents of a *future* diary, it would seem that one's genetic profile *determines*, to a large extent, the person one will become and the life one will lead. Though the predictions may be probabilistic, the suggestion is that discrepancies between what is predicted and what actually happens in the course of one's life represents a *deviation* from what is pre-ordained by the contents of a person's genome.

The connection with autonomy supported by this view is thus deeper than with the property analogy; although we may have property rights *because* we are autonomous, we are only autonomous *if* we are, in fact, able to exercise a meaningful degree of self-determination. If one's future is more or less determined by the contents of one's DNA, however, autonomous choice is illusory. Francis Crick, co-discoverer of the DNA molecule, famously proposed the "astonishing hypothesis" that:

> You, your joys and sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the [genetically determined] behavior of a vast assembly of nerve cells and their associated molecules (Everett 2004, 284).

Crick's hypothesis supports the view that information derived through DNA analysis is distinct from other types of medical information: whereas a person's medical records ordinarily contain information about incidents in one's *past*, the

[29] However, as was discussed in Sect. 2, there are a number of reasons why an analogy with private property is especially problematic in the case of genetic information.

diary analogy suggests genetic information will provide unique insight into one's *future*.[30] The hypothesis suggests that a person's status as self-determining agent is undermined not only by unauthorized access to genetic information, but genetic research itself.

5.2 Rejection of Determinism and Reductionism

Compelling as the metaphors offered by proponents of genetic exceptionalism may be, however, they are more likely to mislead than to enlighten. Crick's "astonishing hypothesis" has been decisively rejected within main-stream genetics (Gilbert and Sarkar 2000; Van Regenmortel 2004). There simply is not the kind of causal connection between a person and her genome posited by proponents of genetic reductionism or, *a fortiori*, genetic determinism. Lewontin (2002, 37) explains:

> The consequence for the understanding of the structure and function of organisms, including their individual and social behavior, is that there is not some small set of universals like Newton's Laws... As is true for living systems in general, relations between genotype and phenotype are contingent, varying from case to case. The outcome of developmental processes depends both on the genotype and on the temporal sequence of environments in which the organism develops.

Early research in clinical genetics focused mainly on the identification of monogenic diseases such as Huntington's disease and Tay-Sachs: these can be identified by locating one point of genetic variation or single nucleotide polymorphism (SNP) within a person's genome. However, most diseases are complex or multi-genic. Complex diseases may be correlated with multiple SNPs and non-genetic factors such as lifestyle and environmental exposures. The shift towards large-scale research thus corresponds with a break in the traditional focus of clinical genetics, namely, the search for one-to-one correspondences between differences in genotype and phenotype. Founti et al. (2009, 578) write:

> The major progress in identifying the genetic basis of Mendelian disorders has not been followed by similar achievements in mapping complex diseases... Inadequate statistical power to detect small and moderate effects was recognized as one of the major limitations. The need for large sample sizes led to numerous large-scale collaborative projects... [and] bio-banks.

Ironically, the failure of genetic reductionism explains both *why* genomic research requires such large data-sets and *how* the right to genetic privacy, in imposing strict and burdensome requirements for individual instances of research access, restricts further progress in genetic research. In the context of bio-banks, the diary analogy is

[30] However the ability to make inferences about a person's future health is not unique to genetic information. High blood pressure, a positive HIV test, or even information about a person's recreational activities all lend similar, and often more significant, insight. See Murray in Rothstein (1999).

particularly misleading because individuals do not participate in genomic research qua individuals, but as members of a large sample group.[31]

New advances in genomic research will rely upon access to large sets of genetic and non-genetic information. For example, a recent study conducted by the International Warfarin Pharmacogenetics Consortium (IWPC) discovered that the required dose of Warfarin[32] depends upon variation in patients' CYP2C9 and VKORC1 genes. The IWPC researchers relied upon pooled bio-bank data, through which "the consortium members had access to anonymized information from about 5700 people on stable dosages of Warfarin … Including Taiwan, Japan, Korea, Singapore, Sweden, Israel, Brazil, Britain and the United States" (National Institute of Health 2009). Jeremy M. Berg, Ph.D., director of the National Institute of General Medical Sciences (NIGMS) which oversaw the study, lauded the practice of international data-sharing and emphasized its importance to future research: "By sharing information and expertise, the consortium researchers developed a way to dose Warfarin that is based on data from patients around the world. This is a highly commendable example of international cooperation and data sharing and should increase the potential utility of the results" (National Institute of Health 2009). However, the access granted to the consortium research groups is explicitly prohibited by the right to genetic privacy; in this case, it would have been prohibitively expensive to re-obtain consent from participants.

5.3 A New Metaphor for Genetic Causation

Recognizing the complex relationship between an organism's environment, genome and biological development, Ridley and Bodmer (1999, 148) offer an alternative to the metaphors employed by genetic reductionism and determinism:

> The brain, the body, and the genome are locked, all three, in a dance. The genome is as much under the control of the other two as they are controlled by it. That is partly why genetic determinism is such a myth. The switching on and off of human genes can be influenced by conscious or unconscious external action.

As this metaphor suggests, the relationship between genetic information and a person's future health is dependent upon factors not revealed through DNA analysis; there are not strict causal connections between genetic variation and personal traits and, contra Crick's hypothesis, the most one can hope for from genomic research are statistical generalizations. However such generalizations have immense utility (Ginsburg and Willard 2009). Knowledge of one's genetic predispositions provides

[31] Ursin notes that this aspect of bio-bank research makes the prominence of personal privacy concerns both misplaced and "paradoxical." See (Ursin 269)

[32] Warfarin, an oral anti-coagulant, is one of the most commonly prescribed drugs in the developed world; it is the second most prescribed dug in America, with over 21 million annual prescriptions. See Abrahams (2010).

valuable information for individuals to reduce risks of developing certain diseases and avoid adverse reactions to medical treatments (Wang et al. 2011). As will be discussed in the following section, this category of information has significant implications for the principle of respect for autonomy.

6 Applications

The previous two sections considered the argument that genetic privacy is intrinsically[33] valuable as a facet of autonomy, and argued that this position is flawed where it posits an untenable connection between privacy and autonomy, and relies upon an empirically false conception of genetic information.

The view that, without explicit and informed consent, genetic research categorically impinges upon autonomy hinges on an impoverished picture of autonomy, one that neglects the fuller picture of the field in which we operate. This view gives ultimate import to the narrow and discrete act of consent; as a result, it forsakes the epistemological horizon of autonomy.[34]

6.1 Genetic Privacy and the Protection of Autonomy

Autonomy depends upon more than the mere absence of external constraint. Consequently, the principle of respect for autonomy should not be interpreted as a blanket duty of non-interference, and does not support an unqualified obligation to respect privacy or obtain informed consent.[35]

Limits on the obligation to obtain informed consent are most obvious in cases involving infants and other populations that either lack the mental capacity to be self-determining or are physically dependent upon the care of others for their survival.[36] For example, the PKU gene is an inherited genetic variation that prevents carriers from metabolizing the amino acid phenylalanine. For children who carry the gene, consuming products that contain phenylalanine, such as artificial sweeteners or high protein foods like milk or meat, will lead to severe mental retardation. Taken literally, the right to genetic privacy would prohibit the practice of newborn

[33] It is still very possible that genetic privacy and autonomy are *consequentially* or *contingently* related, for example if disclosure of genetic information leads to genetic discrimination.

[34] I am indebted to Larry McGrath for this turn of phrase.

[35] For general criticism of the putative connection between informed consent and autonomy, see Manson and O'Neill (2007, Chap. Consent: Nuremberg, Helsinki and beyond).

[36] A number of alternatives to obtaining informed consent from such populations have been proposed. For a discussion of consent and children, see Brock and Buchanan (1990, Chap. 5: Minors).

screening since babies are obviously unable to give informed consent.[37] In the case of PKU, this would mean that the parents of children carrying the gene would not receive crucial information about their child's condition: while the symptoms of PKU are avoidable through simple modifications in a child's diet, there is no cure once the condition is developed.

The case of PKU shows how genetic privacy cannot be intrinsically valuable as a facet of *autonomy*: in this case one would, by respecting genetic privacy, fail to protect development of the cognitive capacities which autonomy requires. Thus one would not only fail to protect a child's welfare, but foreclose her possibility of becoming an autonomous person.

One might object that the case of PKU is exceptional because it involves children who are unable to give informed consent. However, the case also serves as a more general example of how genetic privacy rights may limit public initiatives to detect and treat preventable genetic disorders in individuals (Rothstein 1999, Chap. Genetic Privacy and Public Health). Even where disease does not directly affect an individual's capacity for autonomous choice (e.g. by causing mental impairment), physical disability may hinder an individual in *exercising* his or her autonomy. Autonomous action and choice always presupposes that an individual has the capacity to act and choose. Both capacities can be significantly diminished or destroyed entirely by the onset of preventable genetic disorders. One might add this observation to the list of reasons why respect for autonomy cannot just be interpreted as an obligation of non-interference: autonomous agency requires certain physical and cognitive capacities that can also be limited or destroyed by a *failure* to act (Komrad 1983).

6.2 Genomic Research and the Enhancement of Autonomy

According to Mill, self-determination is valuable precisely because of what is required of a person to *be* autonomous:

> He must use observation to see, reasoning and judgment to foresee, activity to gather materials for decision, discrimination to decide, and when he has decided, firmness and self-control to hold to his deliberate decision. Human nature is not a machine . . . but a tree, which requires to grow and develop itself on all sides, according to the tendency of the inward forces which make it a living thing (Glover 1990, 89).

Beyond minimal physical and cognitive capacities, an agent's ability to choose autonomously depends upon her *awareness* and *understanding* of available options. Thus a person's capacity to make autonomous choices may be diminished not only

[37] Genetic screening programs are under increasing attack from genetic privacy groups in the USA. See Roser (2009).

when options are restricted through *external* interference but also by ignorance in the form of a lack of understanding or unawareness of available choices. In other words, if the value of autonomy consists in the ability of individuals to act in a self-directed manner, autonomous action and knowledge of the conditions in which one operates (e.g. likelihood of certain outcomes, alternatives, etc.) are both practically and conceptually bound up. Practically, we may fail to actually achieve what we set out to accomplish if we act from ignorance; more conceptually, the value of, and obligation to protect, an individual's ability to act on a plan developed from a set of false or incomplete premises is, at best, dubious. Thus, to the extent information provides a fuller picture of the field in which we operate, e.g. by providing a better picture of likely outcomes, sharing or withholding can affect our autonomy.

This is yet another way in which respect for privacy can be distinguished from autonomy. When a person is deceived, privacy is unaffected, but autonomy is diminished; Dworkin (1988, 104) suggests that "such a case involves control over information, but of just the opposite kind at issue in privacy. What is controlled is the information coming to you, not the information coming from you."

In the clinical setting, the principle of respect for autonomy entails inviting patients to participate in their own treatment and prevention decisions (Childress and Beauchamp 2001, 63). However, a patient who must decide whether to start a new course of treatment is only able to decide autonomously, as opposed to arbitrarily, if she understands the likely outcome of her choice: in the absence of such understanding, her choice is not a meaningful expression of her autonomy.

Genomic research enhances autonomy through genomic medicine, "the use of information from genomes and their derivatives (RNA, proteins, and metabolites) to guide medical decision making" (Meyer 2000, 93). The Warfarin study is an example of one recent success in pharmacogenomics, which seeks to identify genetic variations that "can affect an individual's risk of having an adverse drug reaction, or can alter the efficacy of drug treatment in that individual" (Roses 2000, 65). Besides promoting more effective drug prescribing, pharmacogenomics promises to give *patients* a far better understanding of likely side effects that may affect their choice of whether to begin treatment or seek an alternative medication (Wang et al. 2011; Guchelaar et al. 2014). Another field that promises to enhance patient autonomy is perioperative genomics, which studies the relationship between genotypes and surgery outcomes, e.g. success rates and recovery time (Kertai et al. 2015). Pharmacogenomic and perioperative genomic research are not always concerned with developing new treatments *per se*, but rather providing information relevant to treatment *decisions*. Similarly, information about an individual's genetic risks enables that person to make choices that will influence their future health (Gilbert and Sarkar 2000; Podgoreanu and Schwinn 2005; Mick and Faraone 2008; Eitan et al. 2009). The information provided by genomic research gives radical new meaning to the idea of self-determination and, consequently, the principle of respect for autonomy.

7 Conclusion

This chapter began by giving an account of the argument for genetic exceptionalism, and found that the most viable version roots the value of genetic privacy in autonomy. However, following a review of both privacy and autonomy, it became apparent that the argument for genetic exceptionalism misconstrues the connection between genetic information and respect for autonomy. The reality is that genetic research promises to greatly enhance individual autonomy by providing more information about, and control over, health outcomes. In the context of large scale genetic research, respect for autonomy entails implementing systems that not only accommodate individual privacy protections, but also facilitate greater sharing of research data.

Some proponents of genetic privacy rights maintain that there is a need to protect genetic privacy *just because* people consider genetic information private. For example, Rothstein (1999, 459) remarks that genetic exceptionalism is an "overwhelmingly social rather than scientific phenomena. Genetic information is unique because it is regarded as unique" but elsewhere claims that "because genetic information is connected to personal and group identity" genetic privacy is a "fundamental right" (Anderlik and Rothstein 2001, sec. Arguments in favor of protecting privacy). In a similar vein, the authors of the *Genetic Privacy Act* claim not only that genetic information is "powerful and personal" and "uniquely sensitive" (Annas et al. 1995, i) but also that "to the extent that we accord special status to our genes and what they reveal, genetic information . . . merits unique privacy protections" (Roche et al. 1996, 25). In other words, the authors assert that genetic information should be considered private so long as we regard it as private.

This argument is the result of circular reasoning. The authors state their belief that genetic information is exceptional, claim that information believed to be exceptional should be kept private, and conclude that because genetic information is believed to be exceptional, it should be kept private. The argument is not amenable to critical inquiry: it holds that intuitive beliefs, even if they are wrong, are enough to justify genetic privacy as a fundamental right.

However, given the complex nature of genetic information, genetic causation and genomic research, the case for relying upon intuitive notions of privacy is a weak one. One can imagine a world in which phrenology still held sway and, as a consequence, ethicists argued that any information pertaining to the size of one's skull ought to be accorded a special status. Once phrenology was revealed as a pseudo-science, however, we would expect this view to change. The mere fact that some people still believed in phrenology would not, of itself, give justification for treating information about skulls as ethically distinct from other sorts of medical information.

In the present case, it is not only misguided but irresponsible to promote policies that fundamentally misconstrue the nature of both genetic information and genetic research. Policies attempting to protect autonomy through strict controls over genetic information may, ironically, hinder research that, by providing participants

insight into their genetic predispositions, strengthens the capacity for autonomous decision making. If, as this chapter has argued, genetic privacy is not necessarily valuable in its own right, the aim of ethical guidelines should not only be to safeguard against potential abuses of genetic information, but also to ensure that participants and the public at large truly benefit from genetic research. This could, for example, entail a moral imperative to integrate pharmacogenomics into regular medical practice, or reorganize public medical databases such that preventable conditions with genetic correlates can be more readily identified in individuals. The point is that ethical considerations ought to flow in both directions: genetic research is not merely the subject of moral restraints, but also a wellspring for moral obligations.

The author would like to thank Julian Savulescu and Roger Crisp for their supervision during the drafting of this chapter.

References

Abrahams, E. 2010. Latest news & updates from the personalized medicine coalition. *Personalized Medicine* 7(1): 13–14.

Anderlik, M.R., and M.A. Rothstein. 2001. Privacy and confidentiality of genetic information: What rules for the new science? *Annual Review of Genomics and Human Genetics* 2(1): 401–433.

Annas, G.J. 1993. Privacy rules for DNA databanks. Protecting coded 'future diaries'. *Journal of the American Medical Association* 270(19): 2346.

Annas, G.J. 1995. Genetic prophecy and genetic privacy–can we prevent the dream from becoming a nightmare? *American Journal of Public Health* 85(9): 1196.

Annas, G.J. 1999. Genetic privacy: There ought to be a law. *Texas Review of Law and Politics* 4: 9.

Annas, George J., and Michael A. Grodin. 1992. 'The Nazi Doctors and the Nuremberg Code Human Rights in Human Experimentation'. http://philpapers.org/rec/ANNTND.

Annas, G.J, Glantz, L.H., and Roche, P.A. 1995. Drafting the genetic privacy act: Science, policy, and practical considerations. *Journal of Law, Medicine & Ethics* 23: 360.

Arnason, V. 2004. Coding and consent: Moral challenges of the database project in Iceland. *Bioethics* 18: 27–49.

Bathe, O.F., and A.L. McGuire. 2009. The ethical use of existing samples for genome research. *Genetics in Medicine* 11(10): 712–715.

Beskow, L.M., W. Burke, J.F. Merz, P.A. Barr, S. Terry, V.B. Penchaszadeh, L.O. Gostin, M. Gwinn, and M.J. Khoury. 2001. Informed consent for population-based research involving genetics. *Journal of the American Medical Association* 286(18): 2315–2321.

Bove, C.M., S.T. Fry, and D.J. MacDonald. 1997. Presymptomatic and predisposition genetic testing: Ethical and social considerations. *Seminars in Oncology Nursing* 13: 135–140.

Brock, D.W., and Allen E. Buchanan. 1990. *Deciding for others: The ethics of surrogate decision making (studies in philosophy and health policy)*. Cambridge: Cambridge University Press.

Burgess, M.M., S. Adam, M. Bloch, and M.R. Hayden. 1997. Dilemmas of anonymous predictive testing for Huntington disease: Privacy vs. optimal care. *American Journal of Medical Genetics* 71(2): 197–201.

Caplan, A.L. 2009. Economist debates: The ethics of DNA databasing: Statements. Available from http://www.economist.com/debate/days/view/284. Accessed 9 Jan 2010.

Caulfield, T. 2007. Biobanks and blanket consent: The proper place of the public perception and public good rationales. *King's Law Journal* 18: 209–226.
Childress, J.F., and T.L. Beauchamp. 2001. *Principles of biomedical ethics*. New York: Oxford University Press (USA).
Church, G., C. Heeney, N. Hawkins, J. de Vries, P. Boddington, J. Kaye, M. Bobrow, B. Weir, and P3G Consortium. 2009. Public access to genome-wide data: Five views on balancing research with privacy and protection. *Genetics* 5(10): e1000665.
Collins, F.S., and J.D. Watson. 2003. Genetic discrimination: Time to act. *Science* 302(5646): 745.
Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. 1984. Ethics, politics, and access to health care: A critical analysis of the President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. *Cardozo Law Review* 6: 303.
Dickens, B.M., N. Pei, and K.M. Taylor. 1996. Legal and ethical issues in genetic testing and counseling for susceptibility to breast, ovarian and colon cancer. *CMAJ: Canadian Medical Association Journal = Journal De l'Association Medicale Canadienne* 154(6): 813–818.
Diver, C.S., and J.M. Cohen. 2000. Genophobia: What is wrong with genetic discrimination. *University of Pennsylvania Law Review* 149: 1439.
Doukas, David J., and J.W. Berg. 2001. The family covenant and genetic testing'. *The American Journal of Bioethics* 1(3): 2–10. doi:10.1162/152651601750417784.
Dworkin, Ronald. 1978. *Taking rights seriously*. Cambridge: Harvard University Press.
Dworkin, Gerald. 1988. *The theory and practice of autonomy*. Cambridge: Cambridge University Press.
Eitan, Ram, Rachel Michaelson-Cohen, Hanoch Levavi, and Uziel Beller. 2009. The counseling and management of young healthy BRCA mutation carriers. *Journal of Gynecological Cancer October 2009* 19(7): 1156–1159.
Elger, B.S., and A.L. Caplan. 2006. Consent and anonymization in research involving biobanks: Differing terms and norms present serious barriers to an international framework. *EMBO Reports* 7(7): 661–666.
Eriksson, S., and G. Helgesson. 2005. Potential harms, anonymization, and the right to withdraw consent to biobank research. *European Journal of Human Genetics* 13(9): 1071–1076.
Etzioni, A. 2000. *The limits of privacy*. New York: Basic Books.
Everett, M. 2004. Can you keep a (genetic) secret? The genetic privacy movement. *Journal of Genetic Counseling* 13(4): 273–291.
Feinberg, J. 1973. The idea of a free man. In *Educational judgments*, 143–69. London: Routledge.
Founti, P., L. van Koolwijk, C.E. Traverso, N. Pfeiffer, and A.C. Viswanathan. 2009. Biobanks and the importance of detailed phenotyping: A case study—the European Glaucoma Society GlaucoGENE project. *British Journal of Ophthalmology* 93(5): 577–581.
Gilbert, S.F., and S. Sarkar. 2000. Embracing complexity: Organicism for the 21st century. *Developmental Dynamics* 219(1): 1–9.
Ginsburg, G.S., and H. Willard. 2009. Genomic and personalized medicine: Foundations and applications. *Translational Research: The Journal of Laboratory and Clinical Medicine* 154(6): 277–287.
Glover, J. 1990. *Utilitarianism and its critics: Philosophical topics*. New York: Macmillan USA.
Goerl, H.S., R.N. Hyer, and D.H. Farkas. 1997. Genetic privacy legislation: Two views. *Molecular Diagnosis* 2(1): 83–87.
Gostin, L.O. 1995. Genetic privacy. *Journal of Law, Medicine & Ethics* 23: 320.
Greely, H.T. 2007. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annual Review of Genomics and Human Genetics* 8: 343–364.
Green, R.C., D. Lautenbach, and A.L. McGuire. 2015. GINA, genetic discrimination and genomic medicine. *New England Journal of Medicine* 372: 397–399.
Guchelaar, H.-J., H. Gelderblom, T. van der Straaten, J.H.M. Schellens, and J.J. Swen. 2014. Pharmacogenetics in the cancer clinic: From candidate gene studies to next-generation sequencing. *Clinical Pharmacology and Therapeutics* 95(4): 383–385. doi:10.1038/clpt.2014.13.

Gurwitz, D., I. Fortier, J.E. Lunshof, and B.M. Knoppers. 2009. Children and population biobanks. *Science* 325(5942): 818.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339(6117): 321–324.

Hall, M.A., and S.S. Rich. 2000. Patients' fear of genetic discrimination by health insurers: The impact of legal protections. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 2(4): 214–221.

Hansson, M.G. 2009. Ethics and biobanks. *British Journal of Cancer* 100(1): 8–12.

Hansson, M.G., J. Dillner, C.R. Bartram, J.A. Carlsson, and G. Helgesson. 2006. Should donors be allowed to give broad consent to future biobank research? *Lancet Oncology* 7: 266–269.

Homer, N. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *Genetics* 4: e1000167.

Hudson, K. L., et al. 1995. Genetic discrimination and health insurance: An urgent need for reform. *Science* 270: 391–393.

Irwin, T. 1985. *Aristotle: Nicomachean ethics*. Cambridge: Cambridge University Press.

Kaye, J., P. Boddington, J. de Vries, N. Hawkins, and K. Melham. 2009a. Ethical implications of the use of whole genome methods in medical research. *European Journal of Human Genetics* 18(398).

Kaye, J., C. Heeney, N. Hawkins, J. de Vries, and P. Boddington. 2009b. Data sharing in genomics–re-shaping scientific practice. *Nature Reviews Genetics* 10(5): 331–335.

Kertai, Miklos D., Yi-Ju Li, Yen-Wei Li, Yunqi Ji, John Alexander, Mark F. Newman, Peter K. Smith, et al. 2015. Genome-Wide Association Study of perioperative myocardial infarction after coronary artery bypass surgery. *BMJ Open* 5(5), e006920. doi:10.1136/bmjopen-2014-006920.

Komrad, M.S. 1983. A defence of medical paternalism: Maximising patients' autonomy. *Journal of Medical Ethics* 9(1): 38–44.

Konvitz, M.R. 1966. Privacy and the law: A philosophical prelude. *Law & Contemporary Problems* 31: 272.

Krawczak, M., E.V. Ball, I. Fenton, P.D. Stenson, S. Abeysinghe, N. Thomas, and D.N. Cooper. 1999. Human gene mutation database-a biomedical information and research resource. *Human Mutation* 15(1): 45–51.

Kupfer, J. 1987. Privacy, autonomy, and self-concept. *American Philosophical Quarterly* 24(1): 81–89.

Lewontin, R. 2002. *The triple helix: Gene, organism, and environment*. Cambridge: Harvard University Press.

Lunshof, J.E., R. Chadwick, D.B. Vorhaus, and G.M. Church. 2008. From genetic privacy to open consent. *Nature Reviews Genetics* 9(5): 406–410.

Manson, N.C., and O. O'Neill. 2007. *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Margulis, S.T. 2003. Privacy as a social issue and behavioral concept. *Contemporary Perspectives on Privacy: Social, Psychological, Political* 59(2): 243–261.

Mauron, A., and A. Boggio. 2005. Human genetic databases: Towards a global ethical framework. In *La recherche en génétique et en génomique: droits et responsabilités*. Montreal: Les Editions Thémis.

McCloskey, H.J. 1980. Privacy and the right to privacy. *Philosophy* 55(211): 17–38.

McGuire, A.L., R. Fisher, P. Cusenza, K. Hudson, M.A. Rothstein, D. McGraw, S. Matteson, J. Glaser, and D.E. Henley. 2008. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: Points to consider. *Genetics in Medicine* 10(7): 495.

McGuire, A.L., D. Golan, E. Halperin, and Y. Erlich. 2013. Identifying personal genomes by surname inference. *Science (New York, N.Y.)* 339: 321–324. doi:10.1126/science.1229566.

Meyer, U.A. 2000. Pharmacogenetics and adverse drug reactions. *The Lancet* 356(9242): 1667–1671.

Mick, E., and S.V. Faraone. 2008. Genetics of attention deficit hyperactivity disorder. *Child and Adolescent Psychiatric Clinics of North America* 17(2): 261–284.
Mill, J.S. 1869. *On liberty*. London: Longman, Roberts & Green.
Motoc, I. 2009. The international law of genetic discrimination: The power of'never again. In *New technologies and human rights*, ed. Therese Murphy. Oxford: Oxford University Press.
Mulholland, W.F. 1998. Genetic privacy and discrimination: A survey of state legislation. *Jurimetrics* 39: 317.
Nagel, T. 1998. Concealment and exposure. *Philosophy and Public Affairs* 27(1): 3–30.
National Institute of Health. 2009. Could genetics improve Warfarin dosing? Available from http://www.nih.gov/news/health/feb2009/nigms-18.htm. Accessed 30 Dec 2009.
Nevett, L.C. 2001. *House and society in the ancient Greek world*. Cambridge: Cambridge University Press.
Nowlan, W. 2002. Human genetics: A rational view of insurance and genetic discrimination. *Science* 297(5579): 195.
Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
O'Neill, O. 2002. *Autonomy and trust in bioethics (Gifford Lectures, 2001)*. Cambridge: Cambridge University Press.
Oxford English Dictionary. 2010. Privacy, n. Available from http://dictionary.oed.com/cgi/entry/50188914?single=1&query_type=word&queryword=privacy&first=1&max_to_show=10. Accessed14 Feb 2010.
Parent, W.A. 1983a. A new definition of privacy for the law. *Law and Philosophy* 2(3): 305–338.
Parent, W.A. 1983b. Privacy, morality, and the law. *Philosophy & Public Affairs* 12(4): 269–288.
Parthasarathy, S. 2004. Regulating risk: Defining genetic privacy in the United States and Britain. *Science, Technology & Human Values* 29(3): 332–352.
Podgoreanu, M.V., and D.A. Schwinn. 2005. New paradigms in cardiovascular medicine: Emerging technologies and practices: Perioperative genomics. *Journal of the American College of Cardiology* 46(11): 1965–1977.
Rao, R. 2007. Genes and spleens: Property, contract, or privacy rights in the human body. *Journal of Law, Medicine & Ethics* 35: 371.
Ridley, M., and W.F. Bodmer. 1999. *Genome: The autobiography of a species in 23 chapters*. London: Fourth Estate London.
Roche, P., L.H. Glantz, and G.J. Annas. 1996. Genetic Privacy Act: A proposal for national legislation. *The Jurimetrics* 37: 1.
Roser, M.A. 2009. State agrees to destroy more than 5 million stored blood samples from newborns. Available from http://www.statesman.com/news/texas/state-agrees-to-destroy-more-than-5-million-141734.html. Accessed 24 Dec 2009.
Roses, A.D. 2000. Pharmacogenetics and future drug development and delivery. *The Lancet* 355(9212): 1358–1361.
Ross, W.D., and P. Stratton-Lake. 2002. *The right and the good*. New York: Oxford University Press, USA.
Rothstein, M.A. 1999. *Genetic secrets: Protecting privacy and confidentiality in the genetic era*. New Haven: Yale University Press.
Rothstein, M.A. 2005. Genetic exceptionalism & legislative pragmatism. *The Hastings Center Report* 35(4): 27–33.
Rothstein, M.A. 2010. Is deidentification sufficient to protect health privacy in research? *American Journal Bioethics* 10(9): 3–11.
Ryan, A. 1994. Self-ownership, autonomy, and property rights. *Social Philosophy and Policy* 11(2): 241–258.
Salkever, S.G. 1977. Freedom, participation, and happiness. *Political Theory* 5(3): 391–413.
Sankar, P. 2003. Genetic privacy. *Annual Review of Medicine* 54(1): 393–407.
Scanlon, T. 1972. A theory of freedom of expression. *Philosophy & Public Affairs* 1(2): 204–226.
Scanlon, T. 1975. Thomson on privacy. *Philosophy and Public Affairs* 4(4): 315–322.
Sommerville, A., and V. English. 1999. Genetic privacy: Orthodoxy or oxymoron? *Journal of Medical Ethics* 25(2): 144–150.

Steinberg, K.K., M.J. Khoury, E. Thomson, L. Andrews, M.J. Kahn, L.M. Kopelman, J.O. Weiss, and E.W. Clayton. 1995. Informed consent for genetic research on stored tissue samples. *Journal of the American Medical Association* 274(22): 1786–1792.

Suter, S.M. 2001. The Allure and Peril of genetics exceptionalism: Do we need special genetics legislation. *Washington University Law Quarterly* 79: 669.

Suter, S.M. 2003. Disentangling privacy from property: Toward a deeper understanding of genetic privacy. *George Washington Law Review* 72: 737.

Taylor, J.S. 2002. Privacy and autonomy: A reappraisal. *The Southern Journal of Philosophy* 40(4): 587–604.

Taylor, S.D., K.K. Barlow-Stewart, and M.F. Otlowski. 2003. Genetic discrimination: Too few data. *European Journal of Human Genetics* 11(1): 1–2.

Thomson, J.J. 1975. The right to privacy. *Philosophy and Public Affairs* 4(4): 295–314.

United Nations Economic and Social Council. 2004 Genetic privacy and non-discrimination – Report of the Secretary-General. Available from http://www.un.org/ecosoc//docs/report.asp?id=1304. Accessed 5 Jan 2010.

Ursin, L.O. 2008. Biobank research and the right to privacy. *Theoretical Medicine and Bioethics* 29(4): 267–285.

Van Regenmortel, M.H.V. 2004. Biological complexity emerges from the ashes of genetic reductionism. *Journal of Molecular Recognition* 17(3): 145–148.

Visscher, P.M., et al. 2012. Five years of GWAS discovery. *American Journal Human Genetics* 90(1): 7–24.

Wang, Liewei, H.L. McLeod, and R.M. Weinshilboum. 2011. Genomics and drug response. *The New England Journal of Medicine* 364(12): 1144–1153. doi:10.1056/NEJMra1010600.

Warren, S.D., and L.D. Brandeis. 1890. The right to privacy. *Harvard Law Review* 4(5).

Weinreb, L.L. 2009. The right to privacy. *Social Philosophy and Policy* 17(02): 25–44.

Westin, A.E. 2003. Social and political dimensions of privacy. *Contemporary Perspectives on Privacy: Social, Psychological, Political* 59(2): 431–453.

Whitman, J.Q. 2004. The two western cultures of privacy: Dignity versus liberty. *The Yale Law Journal* 113(6): 1151–1221.

Wolff, R.P. 1998. *In defense of anarchism*. Berkeley: University of California Press.

World Medical Association. 2001. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization* 79(4): 373–374.

Part III
Consent

How Data Are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent

J. Patrick Woolley

Abstract The Big Data vision for biomedicine supports compilation, long-term banking, and sharing of sensitive personal information, and potentially allows an individual's data to be combined and utilised with other data indefinitely in innumerable research projects. But this vision does not adhere to the case-specific and jurisdiction-specific oversight models upon which current governance has been founded. Informed consent, for instance, a core principle of bioethics, no longer seems feasible as data repositories and Big Data methodologies become central to research. Policymakers have not yet found a consistent way to address ethical, legal, and social issues (ELSI) in this new data environment. Systematic ways of thinking are needed which reflect the new uses of biomedical data with a view toward upholding basic ELSI standards. The aim of this chapter is to present a view of governance where dataflow itself, not institutional or national boundaries, is taken as the *de facto* framework for research, and where metadata on consent play a central role in how data is governed. I identify types of consent as a place to begin to develop ELSI metadata procedures for data-enabled research contexts. Such metadata can assure data production, dissemination, and reuse is in accordance with participants' and researchers' expectations. Ultimately, it can assist with codification of criteria and standards that demonstrate impact of data intensive research along its ethical, legal, and social dimensions, at multiple levels of governance.

1 Introduction

The power of data collection and analysis is changing how research is done. It has been said Big Data will be as foundational to the next generation of researchers as the internet is to the current generation (Dumbill et al. 2013). Big Data approaches to biomedical research differ fundamentally from other Big Data enterprises, such as finance and marketing, in that historically the data acquired and utilized originate predominantly from biological samples. Those samples and their attendant data

J. Patrick Woolley (✉)
Harris Manchester College, University of Oxford, Mansfield Road, London, OX1 3TD UK
e-mail: drjpatrickwoolley@gmail.com

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_8

have been highly regulated by law and other governance mechanisms put in place to protect research subjects. Ethical and legal protections for medical research participants have thus far been premised on the concepts of informed consent and on the altruistic gifting of samples and personal information for the betterment of society. This model has been supported by the jurisdiction-specific interpretation and implementation of laws, as well as by regional ethical oversight mechanisms, such as expert research ethics committees (RECs and IRBs) and traditional consent forms, designed to provide nuanced consideration of risks associated with the idiosyncrasies of a particular research project.

The Big Data vision for biomedicine, in contrast, supports compilation, long-term banking, and sharing of sensitive personal information, and potentially allows an individual's data to be combined and utilised with other data indefinitely in innumerable research projects. This vision does not adhere to the case-specific and jurisdiction-specific oversight models upon which current governance has been founded. With the advent of Big Data, research is rapidly moving from a one researcher, one project, one jurisdiction model of post-war research, where physical harm was the primary concern (Kaye 2012), to international networks of researchers, where the prevention of informational harm becomes a major challenge.

Public concerns about data sharing are growing. Impending legislation threatens to restrict data access for biomedical research. In their whitepaper, the Global Alliance for Genomics and Health (GA4H), an alliance representing hundreds of institutions in scores of countries seeking to coordinate biomedical data sharing from diverse sources for researchers located around the globe, emphasize we are at a transitional stage where there is a only a short window of opportunity to establish policy for governance that achieves a viable balance of interests among diverse parties, while still maintaining the values of a "civil society" (Altshuler et al. 2013). But the challenges this vision of research poses for policymakers are formidable. Big Data research objectives are dependent upon effective dataflow. That, in turn, is dependent upon a community who trusts to consent to the sharing of personal information.

The aim of this chapter is to present a view of governance where dataflow itself, not institutional or national boundaries, is taken as the *de facto* framework for research, and where metadata[1] related to consent play a central role in how data are governed. First, I highlight basic problems surrounding consent and discuss why the established ethical and legal functions of consent requirements are difficult to maintain within Big Data contexts. Second, I present metadata surrounding consent as a focal point by which to translate ethical and legal functions from traditional research contexts to those of Big Data. I follow with examples of metadata use in related fields.

[1] Here the term "metadata" is used to mean data which convey information surrounding other data.

2 Consent

Historically, "informed consent" came to be a fundamental principle of bioethics, not as a direct result of obscure theoretical deduction, but through the recognition of the real need to protect subjects from abuse by researchers.[2] That governance was coordinated, in part, by adherence to core principles for which consent is central.[3] As the needs of research have changed over the decades, policy has evolved to ensure ethical oversight and medical research maintain a symbiotic relationship (Carlson et al. 2004). Many of these developments – the use of "broad consent" by biobanks for instance – have attempted to accommodate the changing needs of research while still maintaining the integrity of core ethical principles. Now, as we move beyond biobanks to data-enabled science, it is time for policy to evolve once again. How will core ethical principles be translated to Big Data contexts?

2.1 Law, Consent, and Metadata

A basic tension currently exists between the regulation of biomedical information and the goals of research. On the one hand is the protection of basic rights of data providers and fears of abuse by data users. On the other is the promised greater societal good of research into the aetiology of disease.[4] The European Union Data Protection Regulation, now in its developmental stages, seeks to strengthen individual rights surrounding data usage. It is motivated in part by concerns related to data writ large, including social media data, browsing information, and other relatively new sources for information gathering. But it also regulates the use of biomedical data and entails significant changes to consent requirements for participants and to public interest exemptions for attaining consent. These changes promise to have major consequences for how research will be done.

The Wellcome Trust and scores of noncommercial and academic health research organizations recently released a joint position statement (Wellcome Trust 2015)

[2] See: http://www.hhs.gov/ohrp/archive/nurcode.html. See also Spigner (2007).

[3] Respect for autonomy, beneficence, non maleficence, and justice (Beauchamp 2001).

[4] Though they will not be a focus of this chapter, there are cases where this dichotomy does not necessarily apply. Much depends upon where one stands on the issues of privacy and on the technological ability to preserve it in a given research context. It can be argued that protections of privacy can translate into lower risks to participants, and lower risks to participants can tip the scales toward ethical mandates that, since no true harms are incurred by individuals, emphasize the pursuit of public good over and above individual rights. Here, one needs to consider the value of Big Data analytics in cases where they have no clear public benefit (e.g. market research), in cases where analysis is carried out in a manner where risks to privacy are not a real concern, and in cases where they can answer questions in a way that is more protective of privacy than in other methods of research. Each of these scenarios requires a different balancing of individual rights with the public good.

recommending opposition to regulations that impose consent requirements which, they claim, will unduly obstruct research. Specifically, they argue against amendments to Articles 81 and 83 made by the European Parliament which restrict use of personal data in scientific research without specific consent. If these provisions are adopted, they say, "The use of personal data in research without specific consent would be prohibited or become impossible in practice," and "Health and scientific research will be severely threatened." Seeking a better balance between security, public benefits, and rights, they emphasize the need for less restriction in cases where potential societal goods outweigh individual rights, where "an individual's personal data are only used in research when this is proportionate to the potential benefits for society as a whole." They argue the role of research ethics committees are not being recognized as a strong ethical safeguard in protecting individuals' interests,[5] Further, they warn stringent regulations with few exemptions could make broad consent unlawful.[6] This would have serious consequences for models of research and data sharing that currently depend upon the flexibility broad consent allows.

According to this position statement, research should take place only where there is robust and well functioning research ethical oversight. This oversight is what justifies the giving of broad consent to use data as those governing bodies deem best. But is there in fact currently a robust and functional oversight infrastructure for widespread data sharing? Are these necessary governance mechanisms in place yet? At the same time justifications for broad consent are being underwritten by the promise of close ethical oversight, oversight bodies have been identified as major bottlenecks to data-enabled science. Veerus et al. have documented the wide disparity in how ethic committees function and their range of responsibilities from one region to the next, making coordination and harmonization nearly impossible.[7] Lynch et al. document the lack of will by oversight bodies to use legal instruments already available, such as an equivalency principle, which would facilitate international research (Lynch 2014). Dove et al. contrast global disparity in ethical oversight infrastructure, where developing countries have few protections in place, with ethics review infrastructure in industrialized countries, where there are comprehensive protections but they have become politicized, ossified, and dysfunctional (Dove et al. 2013).

Currently, fragmented practices and requirements across jurisdictions make it difficult for research ethics committees (RECs) to establish consistent standards

[5] "The requirement for specific consent fails to take account of the fact that this research is subject to ethical approval and strict confidentiality safeguards, and the identity of individuals is often masked" (Wellcome Trust 2015).

[6] "In many studies that will be affected, individuals have voluntarily given broad consent for their data to be used in research to further our understanding of society, health and disease. Their valuable contributions could be wasted if the amendments become law" (Wellcome Trust 2015).

[7] Lack of uniformity is a major problem (Veerus et al. 2013). GA4GH are looking for ways to streamline oversight and make it more efficient for data use. See: https://genomicsandhealth.org/files/public/3Plenary2Presentation-REWG-BarthaKnoppers-KazutoKato.pdf

for decision-making. Depending upon the data sharing needs of the research, the ethical requirements for RECs to draw together information necessary to make informed decisions can outpace their ability to do so. This makes close oversight of international data sharing appear impossible. In truth, however, much depends upon the type of research in question and the data network through which the data is accessed. Data sharing needs differ significantly in both the regions they cover and the types of research they support. In terms of regulatory concerns, the differences between the UK Biobank and Malaragen, just to take two examples, are vast. Each corresponds to different regulatory requirements for data sharing. The same can be said for CancerGrid, the NCMBI, the Big Data Institute, and countless other data sharing initiatives. And all this is to say nothing of the far more opaque situations for direct to consumer genetic testing and biopharma, where uses (or foreseen potential uses) of proprietary information is held close to the chest.

The position statement above references the European Medical Information Framework (EMIF). Part of its purpose is to address many of the inefficiencies currently limiting the interlinking of heterogeneous datasets across Europe. Ethical oversight is one of the factors they are charged with improving. Also, according to a recent report by the Nuffield Council on Bioethics, *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues* (Nuffield Council on Bioethics 2015) in most cases the perceived risks of data sharing are just that, only perceived risks; actual abuses of data are rare and the risks of data sharing are, at present, low. The costs incurred by research through overly restrictive regulations proposed by the Data Protection Regulation are arguably not justified by the facts. Yet, risks and fears are context dependent, and data sharing does not stop at Europe's shores. Even if good data practices are established for European data, we must also address how research that uses data from sources beyond Europe should be interfaced and regulated. The relatively homogenous regulatory contexts of Europe contrasts with those of the GA4H,[8] where the objective is to share data globally.[9] In international, pan jurisdictional contexts where international law is wanting and there is no EU to enforce regulations, top down strategies can do only so much to preserve rights and protections for participants. This is why many believe allowing participants to give broad consent is vital.

However, broad consent has itself come under fire. Its ethical status has been a source of serious debate in its own right (Caulfield 2009). The question is whether or not it constitutes informed consent. Once, it was seen as an imperfect but necessary stopgap for addressing the consent challenges of biobanks. But now some argue it does in fact satisfy the high ethical standard set by informed consent (Sheehan 2011a, b). Further, some point out *not* allowing broad consent and its broad berth for ethics oversight violates the wishes of many willing participant who want to share their data widely (Dove et al. 2013). Wherever one falls on the issue broad consent in principle, the pragmatic repercussions to its use will be difficult to avoid. If broad

[8] Several signators are from the Wellcome Trust.

[9] See: http://genomicsandhealth.org/about-global-alliance

consent is made impossible by the Data Protection Regulation, many current models for research and data sharing that now depend upon it will have to be changed. And, even if broad consent does survive legislation, proof of well functioning and robust ethical oversight that administers data along the lines of broad consent is still needed to retain public trust. Again, public trust is an issue. Even if it is true that resistance to data sharing is due to reactionary fears to perceived risks, not real ones, it is very difficult to allay fears at the same time individual rights are being taken away by relaxing consent requirements.

Fortunately, there are new ICT supported approaches to consent that offer solutions to many of these problems. UK industry and academia have collaborated on Ensuring Consent and Revocation (EnCoRe) and Registries for All (Reg4ALL). Partnering with academic centres such as the Centre for Health, Law and Emerging Technologies at Oxford (HeLEX), these have developed and tested dynamic consent as an interactive interface that allows participants in research to choose and alter consent choices in real time. The system provides reliable storage and enforcement of these choices by cryptographically protecting sensitive personal information in a way that allows data to be accessed in only those ways for which consented has been given (Kaye et al. 2014). It preserves an option for broad consent. And, if broad consent is ultimately made illegal, it can tailor consent options to conform to new legal conditions. Most importantly, dynamic consent allays participants' fears by putting them in control or their data. These incisive solutions for the problem of consent in data-enabled research contexts will be increasingly important as data sharing methods, and public opinion of it, develops.

However, significant hurdles must be overcome before ICT enabled dynamic consent can be widely implemented. Besides time, money, effort, and the technical capacity to interface with the system, deployment of dynamic consent would require partnerships that "understand and value the central role that patients have in research as the providers of information and biological material." This, in turn, requires new policies, standards, and practices that support this approach. Commitment to this vision must be made across the board, by clinicians, researchers, health-care services, research institutions, and governments. This raises a real problem. Consent requirements imposed by the Data Protection Regulation, or similar legislation, may be enacted far sooner than dynamic consent can be implemented on scale wide enough to allow research to continue uninterrupted. If the transition is not a smooth one, public trust in research could degrade further, restrictions could increase, and research funding could suffer. If dynamic consent does not overcome these hurdles soon enough, its potential might never be realized.

Metadata on consent offer a way forward. The approach to metadata proposed here would not be as robust and functional as widespread implementation of dynamic consent. But it can more immediately be implemented through protocols and best practices where resources for dynamic consent are not available. The objective is to translate the function consent plays for RECs in biobank contexts to data-enabled contexts. And, looking beyond the internal governance structures developed for individual biobanks, this translation should harmonize with different levels and types of governance mechanisms, such as data access committees or

regional committees that approve research projects, which may each have their own ethical remits concerning data sharing. This can be accomplished through consent matrices that provide metadata on conditions surrounding consent.[10]

Such metadata can provide valuable information for ethical oversight bodies which allows them to respond to all types of consent appropriately. Availability of the right kinds of metadata offers a bottom up approach to governance that can assist in a number of ways. First, access to the same kinds of metadata can support greater uniformity among RECs and international oversight bodies in a way that is responsive to the data themselves. This is more effective than top down regulations which may not be as sensitive to the particular circumstances surrounding the data or research. Second, making oversight more functional through metadata could help to allay public fears, and perhaps strengthen arguments for the viability of broad consent long enough to phase in a dynamic consent model. Third, adopting these practices can help to focus, harmonize, and orchestrate communitywide ELSI efforts, ideally leading to eventual widespread adoption of dynamic consent. If dynamic consent is ultimately proven the best solution, these metadata procedures can help to pave the way for its widespread adoption.

Before I discuss strategies for metadata surrounding consent, we must look at the reciprocal relationship between consent requirements and research ethics committees. Appreciating this relationship helps to make it clear why transferring broad consent from biobank research to data-enabled research is not as straightforward as it may first appear.

2.2 The Transfer of Burden from the Participant to Ethics Committees

As research evolves from a traditional project-specific and jurisdiction-specific model, to a biobank model, to the ICT data intensive model that makes Big Data methodologies possible, the pragmatic limitations of consent are being realised. There is no consensus which principles should be supported. Some champion the rights of research participants to remain autonomous and in control of their data[11]; others champion the benefits of science for society and work to keep data usage in the hands of the scientists and ethics committees, as long as participants agree to this

[10]The GA4GH has recently completed a study that examines variation in data use conditions which are based on consent provisions for genomics datasets in both research and clinical settings. The study reviews guidance of the National Institutes of Health, data use conditions at Broad Institute (of MIT and Harvard) and the European Genome-phenome Archive of the European Bioinformatics Institute (EMBL-EBI), as well as data use conditions within the GA4GH's own Data Working Group and the Matchmaker Exchange Project. Their proposed structure for recording categories and requirements for data and data use are in many ways compatible with metadata objectives proposed in this chapter. (Dyke et al. 2016).

[11]For an example see Kaye et al. (2014).

arrangement.[12] These differences are not immediately clear, however, because often different visions are supported by quite similar language on consent. One needs to look beyond the surface to see what ethical perspective is being advocated.

A paper by Steinsbekk et al. on the difference between broad consent and dynamic consent illustrates how tensions between deontological (individual rights) and consequentialist (societal goods) perspectives come to bear on the issue of consent. They present dynamic consent and broad consent to be at odds (Steinsbekk et al. 2013). Yet, as mentioned above, options offered by dynamic consent can include broad consent. So, if the two are not in fact mutually exclusive, why are they presented as if they were in opposition to one another? There is a deeper issue in play. Steinsbekk et al. make the claim it is more ethical to delimit the choices of those who agree to participate in longitudinal biobank research than to present them with the range of options dynamic consent could offer. They say presenting participants with the information necessary to allow them to make informed decisions about how their samples and data are used runs the risk of "individualizing the ethical review of research projects." Putting participants in charge "raises the guard" of participants instead of having them place trust in research ethics committees and researchers. They say, "For most people, we suspect that biomedical research is complex, complicated and rather boring stuff," and the problem with dynamics consent is "very few participants will probably be able to meet [the] high expectations" placed upon them to make informed decisons. They argue having more options will ultimately lessen the availability of data and decrease ethical oversight. In the end, the potential societal benefits of access to data outweigh ethical concerns about informed, autonomous decision making.[13]

The attitude expressed here is a decidedly paternalistic one.[14] Broad consent should be kept, dynamic consent rejected, because limiting the range of participants' options is more ethical than leaving decisions about how their data should be used to them. The tension expressed has less to do with the deontological legitimacy of dynamic consent and more to do with maximising researchers' access to data, allowing health research to proceed with as few encumbrances as possible. In making this case, the authors take a significant consequentialist turn that departs from the deontological framework through which so much law and governance has been framed since the first article of the Nuremberg Code laid out the requirement for consent.

[12]Though agreement may sometimes be attained through "presumed consent," an "opt out" model. See for example: http://www.nhs.uk/NHSEngland/thenhs/records/healthrecords/Pages/care-data.aspx. See also, McCartney (2014).

[13]"Biomedical research, however, is not primarily about our own health but rather about potential health benefit for future generations. An important reason for active engagement and participation in biomedical research is thereby lacking compared with general health care" (Steinsbekk et al. 2013).

[14]Informed consent is also sometimes criticised as being paternalistic in that it can prevent the giving of consent, prevent the exercise of autonomy, if requirements for what "informed" means are not met.

This example illustrates that there can be deeper normative concerns lurking beneath arguments for one or another form of consent. One needs to consider carefully whether or not the rationale guiding a particular argument is in keeping with principles of established ethics and law which we may want to preserve. This is one of the questions being grappled with as policymakers seek to accommodate Big Data biomedical research; the principle of consent has been modified to suit so many different contexts and goals of research, trying to pinpoint its exact ethical function in the literature has become like trying to hit moving target.[15] Some are worried this moving target will inevitably result in a slight of hand, where one form of consent is substituted for another in a way that gives the impression of continuity of well established ethical principles when in fact there is none (Karlsen et al. 2011). This, arguably, is what is occurring above in the debate over dynamic consent. Here we see a shift from a participant-centric, deontological approach to a consequentialist, more paternalistic one, all while maintaining a certain consistency of consent language.[16] We should not put too much emphasis on terminology alone, without analyzing the underlying principles being advocated. This runs the risk of, not only losing the theoretical threads that bind principles of governance together, but also of reducing the concept of consent to "a hollow repetitive ritual, devoid of any relevant moral content" (Karlsen et al. 2011).

In Big Data contexts, subtle shifts in ways language is used could lead to very real pragmatic problems. We see how problems crop up, for example, by examining a debate which considers whether broad consent satisfies the requirements of informed consent for biobanks. In the past several decades, biobanks have been driving a major transition in the way policy surrounding biomedical research is conceived. Biobanks are entrusted with stewardship of, not only biosamples, but also of their attendant data. Depending on the type of biobank, the data can range from basic information on those from whom the samples are taken, to entire genome sequences of sample cells. This establishes a certain continuum as we move from traditional bench top research, to biobanks, to sharing data from those biobanks, to sharing data in general for Big Data research. It is tempting to think, because there is a continuum in the development of research, there is also a continuum in regulation needs. Insofar as a given research project depends upon the examination of extracted and transferable data, and not on the examination of actual biospecimens, biobank policies surrounding data do appear to help close the gap between traditional models of research and those needed for Big Data research.

But this apparent parallel can be misleading. It should not be assumed the governance solutions for the former are transferable to the latter. Solutions for one context do not necessarily offer solutions for the other. We see why this is so, for instance, in Sheehan's argument for broad consent. In contrast to the arguments

[15] In addition to "informed consent," "broad consent," and "dynamic consent," we have "tiered consent," "blanket consent," "open consent," "presumed consent," "implied consent," "precautionary consent," and "waiver of consent" (in retrospective research).

[16] This has been likened to Wittgenstein's language-game. (Karlsen et al. 2011).

made above where subtle shifts in consent language can conceal seismic shifts in thinking on ethical oversight, Sheehan addresses the issue of informed consent head-on. He argues broad consent does not indicate a compromising of informed consent, but merely a new way of thinking about it where there is a transfer of normative responsibility from participants to research ethics committees. This argument assumes a complementary relationship between the forms of consent required by biobanks and the function of the ethics committee within a biobank's institutional architecture. Yet, even if this argument is accepted, this relationship does not readily translate to ICT contexts. Here, the validity of broad consent depends upon a clearly defined and effective research ethics committee, and yet, as we saw above, research ethics committees are precisely what are coming under fire in ICT contexts, criticized for being incapable of overseeing data-enabled science. If we try to transplant Sheehan's argument made for broad consent in biobanks to Big Data objectives, the relationship between consent and research ethics committees comes uncoupled, and the case for broad consent fails.

Let us unpack this illustration further. Informed consent requires a participant is informed of risks and benefits of research. Only then is a potential participant able to exercise his or her autonomy and decide whether or not to assume whatever burdens the research may impose. These need to be derived from real world contexts and scenarios. But, as research is more open-ended in the case of biobanks, risks and benefits cannot be determined prior to research. It is precisely where knowledge of real world scenarios is limited that broad consent is required. Ethical factors that cannot be ascertained at the time of consent are addressed later down the line by research ethics committees. It is left to them to ensure the giving of broad consent does not lead to unacceptably harmful usage of samples or data. The exercise of autonomy by a participant, then, is not in assuming immediate burdens outright, but in putting trust in the research ethics committee to mange future burdens soundly. The ethical veracity of broad consent therefore depends upon there being fully functioning, highly effective research ethics committees who represent the interests of participants as stewards of their data.[17] Trust in them is what makes the process work. It is the trust that participants have in these committees which compels them to give broad consent, even when many factors surrounding research are unknown. And, it is the effectiveness of research ethics committee that makes this arrangement ethically sound.

However, even if Sheehan's argument for broad consent were accepted for biobanks that demonstrate a robust and effective research ethics committee system is in place, the argument does not necessarily translate well into Big Data contexts where datasets from biobanks are combined with those from sources originating beyond their immediate purview. As data from different biobanks are shared and

[17] "[T]he information that makes them informed is different from the specific individual consent case. In broad consent cases, the relevant information is about the person (or institution) who will make the decision for me. In biobanking, the relevant information might be about the overall goals of the research supported by the biobank and details of the decision making processes within the institution – how are decisions made about suitable research and by whom?" (Sheehan 2011a).

applied toward common research objectives, in principle, so too are the remits of the respective data sources. To assume that these remits align or overlap well enough to still justify Sheehan's argument that broad consent constitutes informed consent is to assume that a high degree of transitivity already exists between biobanks' missions, that a very substantial and precise harmonization of missions and standards already exists. But, if this were the case, why do we see the resistance of RECs to share data? (See above.) If this unwillingness to share is due to a conflict with missions, then the data should not be shared. Broad consent does not here support Big Data objectives because it does not adequately free the data for research. Alternatively, if missions are in fact aligned and this unwillingness to share it is due to REC dysfunction or lack of ability to coordinate, then the RECs simply prove themselves unable to carry out what would otherwise be effective data sharing practices. Broad consent does not here support Big Data objectives because the argument for broad consent is based upon highly functioning RECs which can recognise commonalities of missions among themselves, and we see dysfunction instead. The former case is a misalignment of mission principles, the latter case is a misalignment of procedures, logistics, and pragmatics, but in both cases the effect is the same: data sharing practices that would support Big Data objectives are suppressed.

So, while it may at first seem broad consent is what is needed for Big Data objectives, when we look more deeply we find a contradiction. In an attempt free up consent requirements, some legitimize the move away from traditional informed consent to broad consent by putting more and more of the onus of decision-making on research ethics committees. But, at the same time, others are finding research ethics committees to be a serious obstacle for data-enabled science. Ethical oversight bodies have come under fire for being incapable of governing data driven biomedical research. It has been said they are in need of major reform in how data is managed.[18] Arguments for broad consent as informed consent in Big Data research are thus locked into an untenable position: one cannot say the function that RECs perform offers us the solution to ethical problems concerning informed consent while, at the same time, the function that RECs perform is identified as one of the major obstacles to data sharing.

All this raises the question whether specialised forms of ethics committees, or Data Access Committees, designed specifically for data-enabled research can be established and given the tools they need to oversee Big Data research. Below, I propose a way metadata can help to provide these tools.

[18]For examples of calls for institutional reform see: Altshuler et al. (2013), Dove et al. (2013), and Li Ka Shing et al. (2012).

3 Metadata

The arguments above are intended to make a case for why respect for the principle of autonomy, as expressed though consent, should be a part of ethical governance of data in data-enabled research, just as it has been in the ethical governance of participants in clinical trials, and of samples in biobanks. We have taken the deontological, rights-based, participant-centric view on consent to be the default. It is one enshrined in traditional models of ethics and law. The burden has been placed on those who wish to argue for a consequentialist view (or another perspective) to demonstrate that an abandonment of this deontological stance does not unduly undermine the ethical and legal functions of autonomy.[19] It has been argued broad consent developed for biobanks is not necessarily a sound model for the sharing of data in Big Data biomedical research. This is because, even if broad consent were accepted as ethically sound for biobanks because it does in fact constitute a type of informed consent (as per Sheehan), the role of research ethics committees necessary to support this view cannot currently be maintained in ICT contexts where there is a high degree of data sharing and interlinking of data sets. As discussed above, to be viable, broad consent requires a highly functioning research ethics committee that specialises in the specifics of a given biobank's mission. It is this mission and this specificity of oversight to which a participant is consenting. Because the unknowns surrounding broad consent are much more pronounced than in traditional informed consent, an extra burden is transferred to research ethics committees who then determine how the sample and data are to be used. However, the dysfunction of these committees is precisely what is coming under fire as the paradigm shifts from traditional research toward data-enabled science; a reciprocal relationship between consent and responsive research ethics committees is exactly what is lacking in Big Data contexts. In absence of this rationale for broad consent, it is unclear what ethical remit is governing decisions to share data, and unclear how or whether autonomy of participants is being respected. If a basic respect for autonomy is not clear, research can become suspect for violating basic principles of bioethics, as well as basic rights of individuals. We thus run the risk of degrading public trust, and fail to prevent foreseeable public backlashes to data sharing and the creation if regulatory measures unfavourable for research into the aetiology of disease. If the goal is to retain trust by protecting research participants' interests, asking participants what those interests are is a necessary first step. This is communicated through the giving of consent, and ensuring this occurs is in fact what "respect for autonomy" means. If consent is to retain this established ethical function, how can research ethics committees and other governance bodies be made effective in Big Data contexts? Or, if data-enabled research requires different models for oversight, how will those governing bodies and mechanisms function in ICT environments?

[19] As per Steinsbekk's argument above.

For normative problems arising in data contexts, normatively informative metadata should be seen as part of the solution.

3.1 Responding to the Shift in Paradigm

It is important to characterize the challenges properly. To determine what "Big Data" means for policy on biomedical research, determining what "data" means in a given project is a necessary first step. But different research objectives require different methodologies. Different methodologies require different data sources. Different data sources require populations who are differently distributed, often entailing differing jurisdictional issues, and eliciting differing types of ethical and legal concerns. In the end, the simple, blanket term "Big Data" obscures more than it reveals when it comes to guiding policy.

Perhaps a more helpful phrase for focussing thinking on policy development is "dataflow." While the primary challenge for biomedical researchers is having access to the very large quantities of data being produced, policymakers and ELSI researchers need to focus on the origins and histories of that data as it is produced and used. The *de facto* frameworks for policy development are no longer research projects, biobanks, institutions, nor national boundaries, but data infrastructures themselves. When we look beyond the data to dataflow, we see digital data management practices are changing the ways information is produced, stored, and disseminated. Data sharing federations are forming, and data are being integrated from heterogeneous sources to create globally accessible repositories. Increasingly, data distributed across multiple types of information systems designed for particular tasks and purposes are being shared, transferred, and repurposed. Data are being coordinated, networked, and made available for research purposes different than those for which the data were initially collected.

What has traditionally been the role of policy here? Before biomedical data are originally obtained, researchers follow regulatory requirements, ethics committee recommendations, and best practices guidelines as research subjects are chosen, informed of the research project's purposes and risks, and (usually) asked to give consent to having their samples and data used. What protocols are followed depends upon type of research, purpose of research, locations of participants and researchers, types of samples and data taken, and so forth. However, due to the rapid rate of development, it has been difficult for policymakers to keep up with the way data-enabled science is done. Though abuses have not yet been shown to be a major problem (Nuffield Council on Bioethics 2015). preserving the information and contexts of important ethical value in safeguarding against harmful uses of data has thus far not been made part of the process. When data are stripped of the contexts in which they were gathered, reuse of data can potentially go against the wishes of research participants, or contradict original agreements between researchers and research subjects (de Vries et al. 2014).

Developing requirements for metadata could restore this information. Much depends upon the type of research in question and the data network through which

the data is accessed. The degree of usefulness of metadata depends upon the circumstances under which data is shared. A highly monitored and restricted portal that controls and records who has had access to what data and for what purposes is not likely to have an immediate use for metadata if it essentially records the same information. On the other hand, in cases where data collection, curation, and sharing is maximised, such as with the European Molecular Biology Laboratory (EMBL-EBI), the metadata could provide a way to ensure the necessary conditions surrounding consent are part of the decision making process to share data. Also, metadata standards as a prerequisite for data sharing in international contexts could help to minimise abuse. For instance, they could impose standards for researchers who use data originating from countries which do not have the infrastructure and education necessary to support more advanced consent procedures. ELSI related metadata beyond just consent could help in this global arena of data sharing. For instance, information that helps to identify conditions under which data is attained could be used to determine the eligibility of datasets form other countries where the principle of consent is not a central part of legal requirements or cultural value systems.[20] This allows one to determine whether its use remains legal in the West as regional and international laws evolves.

Yet, the ways data are currently shared strips it of much of the information relevant for ELSI concerns (de Vries et al. 2014). Information a well-functioning research ethics committee would need to make evaluations is lost as data are aggregated, managed, and shared. Metadata surrounding consent is an essential component here. They provide the information needed to evaluate data usage along ethical, legal, and social lines. Requiring consistent metadata standards would help to coordinate and harmonize research, oversight, and efforts of the greater ELSI community. It would pave the way for implementation of more advanced consent procedures appropriate for ICT contexts, such as dynamic consent.

3.2 Preliminaries for Metadata on Consent

Metadata methods need to be developed so data which originates at the project or institutional level carries associated information that can be utilized by ELSI researches as the data are collected, transferred, and archived. It is the ELSI community, not scientists or ICT specialists, who are in position to identify what metadata are meaningful for ethical, legal, and social issues. It is their purview, and their responsibility to decide what information is useful and what is not. Ideally, this will occur as biological data infrastructures are still in their early

[20]For instance in South Africa ownership of genomic information can be tribal. Consent is thereby given by tribal leaders, not individuals. This is a consent model that challenges autonomy as it is conceived in Western law and bioethics, yet it is one that respects and is responsive to cultural needs surrounding international data sharing.

stages of development so they can be specifically designed to support the needs of participants, policymakers, and the wider ELSI community.

3.2.1 Consent as a Focal Point: The Consent Matrix

Making available metadata relevant for ELSI concerns requires ELSI researchers, policymakers, and regulators to work closely with biomedical researchers and ICT specialists. But progress is stymied, in part, by an absence of common platforms upon which concerns can be systematically addressed. Here it is suggested, a midst the dizzying complexity of factors involved in determining what ethical uses of biomedical data are, the principle of consent provides a useful focal point, and a consent matrix provides a usefultool.

While the tasks and responsibilities of research ethic committees resist algorithmic thinking, the autonomous giving of consent or not does approach a "Boolean" principle[21] that penetrates to the core of ethical and legal systems of the West. This pragmatic aspect of consent is one of the more promising places to begin to develop metadata standards for the ethical, legal, and social consequences of biomedical research undertaken in ICT contexts.

Consent is a principle consistently recorded by technical forms and procedures that delineate many of the conditions surrounding research: When the data was collected what were the conditions surrounding the consent to it being used? What level of risk was agreed to? What values were expected to be maintained? As such, it offers a means by which to systematically aggregate metadata on various factors that ELSI researchers need for policy development. This metadata can facilitate e-governance by allowing records in diverse locations which employ their own standards for consent to be made interoperable by developing standardized representation, controlled vocabularies, and common definitions that allow semantic integration across collections (Gartner 2013; Kaye 2011).

A consent matrix is a method for gathering this information. The idea of a consent matrix has been implemented in a number of ways in the field of biomedicine and beyond (Dimitropoulos 2013; Fléchais 2005; Katina et al. 2010; Linzer 1988; Tan 2002; Willison 2003). Dynamic consent utilises an elaborate consent matrix. Simpler, web based consent matrices have been piloted (Thiel et al. 2015). A matrix can capture the conditions agreed to on a consent form, and make them available for later use. But, unlike a typical consent form where ticking "yes" is often the only option if consent is to be given, a consent matrix is capable of capturing more information. Through a normatively appropriate choice architecture, it can provide a "Boolean" yes/no format which makes what is not agreed to as much a factor as what is agreed to.[22]

[21] As compared to "beneficence," for example.

[22] I use the term "Boolean" for this example here. But in fact the variables can have more than two options. They can use as many as are needed.

In terms of patient interface, a consent matrix is analogous to a traditional consent form. It can be incorporated into everyday data collection through analogous procedures. A consent matrix does not replace a consent form, but can be integrated into or appended onto one.[23] Anytime a consent form is required, so too can a consent matrix be required. Its form can be paper and pen, digital, or web based. It can be designed to be for "broad consent," "assumed consent," "implied consent," "informed consent," and so on.

Its purpose is not to perform a normative function, but to gather information necessary to make informed normative analysis possible later. As the consent matrix is filled out, for instance if a participant is asked to tick "yes" or "no" in the applicable fields, this determines a specific permutation of the variables. Even if the numbers of variables in this matrix becomes quite large, the number of possible permutations will always be finite. Outcomes can be rerecorded with a simple alphanumeric reference. There are many way to do this. Each option in the matrix may be assigned a place in a reference number, or all possible permutations could be assigned a unique code. An individual's data can be tagged with this code so that the conditions surrounding consent for a given participant can be known by an oversight body. That tag would attend the data and be available for other oversight bodies throughout dataflow. Information on conditions surrounding consent would thereby be translated from the localised format in which they originated, into a format which would allow access in globally distributed, nonlocalised, dataenabled research models. A universal consent matrix could gradually be developed by consolidating as many variables as is required for different kinds of research. In this way, a single consent matrix can in principle consolidate all the variables covered across the many forms of consent now in use. The variables of this matrix can start with the basics and increase with time to capture as much granularity of information as is deemed necessary.

This use of a consent matrix differs from dynamic consent discussed above in that the consent matrix is not updated by the participant and there is no conduit of communication established between the participant and the researchers who use and reuse the data. It does not, therefore solve the problem of recontact, an issue the Wellcome Trust say will make research impossible if proposed amendments to the European Data Protection Regulation goes through. But, on the other hand, metadata a matrix could provide would improve data stewards' ability to be responsive to participants' wishes—improving the performance of even well-funded and highly coordinated oversight bodies such as the EMIF—and thus make legislation unfavorable for data sharing less likely in the future. Requiring use of consent matrices would establish a degree of process uniformity across all data collection contexts which is easily integrated into the day to day workings of researchers and members of the ELSI community. Adopting the process would, in itself, promote harmonization of practices across the biomedical research landscape. Focussing

[23] It could, for instance, be little different than a survey that attends the consent form, though its function is very different.

biomedical research consent requirements around a consent matrix provides a clear, nondisruptive, and nonrestrictive focal point which allows policymakers and the ELSI community to orchestrate efforts and potentially harmonize procedures on a global scale. It permits research and data sharing to go forward along the trajectories currently planned, so efforts here are not lost.

Though this process falls short of the dynamic consent ideal, it does help to overcome the barriers discussed above that now stand in the way of its widespread implementation. It does this by harmonising communitywide procedures along the lines of consent matrices, a tool central to dynamic consent, thus creating the conditions needed for the ready adoption of its ICT based process. This creates a method for bringing about the "cultural change" dynamic consent requires, and makes conditions favourable for dynamic consent to be integrated as economic, technological, and cultural conditions allow.

3.2.2 Data Organizing Behaviours

The way to determine what variables are needed for the consent matrix, both in terms of baseline minimums, and in term of high granularity, is to study how such information is managed now. Metadata production is not about rules based decision making, but about rules based information gathering, not unlike those already in place in the development of consent forms and research ethics committee application systems. Metadata procedures would not replace these systems. It would make information available to augment and enhance ability of oversight bodies to better evaluate data intensive research projects.

Studies have been conducted on data organizing behaviours of information managers and scientists for the purpose of developing ways to make disparate data practices interoperable through metadata strategies for data repositories (Willis et al. 2012). Similar studies are needed on those who organize and manage information for ELSI objectives. To determine what kind of information is useful and how it is used, process mapping is needed on current requirements for how information surrounding biomedical data is currently managed. This can be done for multiple types of biomedical research projects which require multiple forms of data and multiple tiers of governance: local,[24] national,[25] and global research models.[26]

[24] See: http://www.hra.nhs.uk/research-community/applying-for-approvals/

[25] See for examples: UK Biobank consent information at http://www.ukbiobank.ac.uk/wp-content/uploads/2011/06/Participant_information_leaflet.pdf?phpMyAdmin=trmKQlYdjjnQIgJ%2CfAzikMhEnx6; Genomics England, 100,000 Genomes Project consent form at http://www.genomicsengland.co.uk/wp-content/uploads/2015/02/3b_CFProbandAdultPatientor-their-AdultRelsRareDisease-v2.0.pdf

[26] See for examples: GA4GH consent tools at http://www.p3g.org/news/consent-tools-prepared-global-alliance-genomics-and-health-p3g-ipac; International Cancer Genome Consortium (ICGC) Research Study Model Consent Brochure https://icgc.org/files/daco/ICGC_prosp_consent_290110.pdf. See also Wallace (2011).

These processes will differ between projects, regions, and institutions but, insofar as they adhere to them, they will ultimately map to basic functions of more broadly established law and governance.

Attention to many of these factors is already a standard requirement for data collection. But procedures for making this information part of a metadata record that attends data at each stage of use have yet to be developed. In some cases, it may simply be a matter of reverse engineering procedures that have already been developed in analogue form. For example, much can be borrowed from the decision making process that goes into developing 'paper and pen' consent forms, or applications for review processes, many of which have already been translated into rules based website interfaces which, by means of a 'decision tree' design, create datasets to describe specific conditions surrounding research and attach them to research applications.[27] Developing metadata procedures that uniformly record the necessary information at the time it is available, links it to the data, and ensures it is accessible over the lifespan of the data would streamline the research approval process for research that employs secondary uses of data.

3.2.3 Defining Data Elements: Variables of a Consent Matrix

A benefit of consent matrices is that they help to identify and clearly define data elements relevant for ELSI concerns. The more ELSI researchers and policymakers can think in terms of ICT contexts, the more this potential can be realised. A controlled vocabulary is not essential. In principle, a consent matrix can take on an ever-increasing number of options to accommodate any number of possible circumstances. However, the more consistency there is in vocabulary, the more easily redundancies and equivalencies can be spotted. The more controlled the vocabularies are, the more precise and concise the consent matrix can become. To accomplish this, data-related literacy and skills for managing data rich research contexts should become part of ELSI training. Opportunities should be identified for increasing literacy on the roles of metadata in digital data management practices (Qin 2010). This has three related aims. First, it can inform policymakers on various ways information professional and scientists organize data sets for personal use and repository collections (Willis et al. 2012). Second, it can demonstrate to policymakers how metadata allow semantic integration of records in diverse locations that employ differing standards and vocabularies (Gartner 2013). Third, this knowledge can help to translate established principles and categories in ELSI sub disciplines into controlled vocabularies for metadata purposes.

A consent matrix can be expanded to include multiple parts. For instance, a part filled out by the participant can convey the participants wishes, another part filled out by designated oversight bodies can capture conditions around which consent was given, and another part filled out by REC's or researchers can capture

[27] See: http://www.hra.nhs.uk/research-community/applying-for-approvals/

history of data usage.[28] The basic objective is to identify descriptors that can capture ethical, legal, and social factors surrounding biomedical research from its inception, to data production, through to the entire history of data usage. In addition to research participants' concerns, these factors can include conditions surrounding sample collection, funding sources, data transaction histories, and so forth. This requires isolating information on conditions that guide decision-making, identifying data elements that reflect those attributes, and incorporating these elements into wider metadata efforts now taking place in many corners of research and governance. Ultimately, it assists with transparency, oversight, and protection through codification of criteria and standards that demonstrate impact of data intensive research along its ethical, legal, and social dimensions, at multiple levels of governance.

The elements should be nonnormative information, but should allow those who access the metadata to make normative judgments about research that utilizes the data it describes.[29] Having this kind of information available facilitates selection of data based on conditions surrounding it, allowing experimental design to be coordinated with ELSI concerns before the data are used in a research project (Kaye 2011). It allows ready access to ethical. legal, and social issues surrounding research conditions[30] to assure data production, dissemination, and reuse is in accordance with both participants' and researchers' expectations, for instance in accordance with conditions surrounding consent. In short, it would aid in translating the traditional roles of research ethics committees into ICT contexts.

Once data elements are identified, it needs to be determined if greater uniformity may be established across platforms. This task may not be as daunting as it first appears. Work has been done in other fields that shows disparity in data production and storage practices does not reflect disparity in metadata goals; much more uniformity exists across platforms than might be expected. This uniformity facilitates coordination around common objectives that support data sharing in general (Willis et al. 2012). A similar study is needed on the ELSI front. It can be augmented by empirical studies conducted on participants' views on data sharing which detail specific types of information participants' want to know before agree-

[28]This latter part would not have to be part of the matrix. Ideally it would, but it would require that the matrix code is updated every time data us used. Integrating updating procedures into the process is a greater challenge than simply capturing the initial conditions of data collection. It would require a detailed knowledge of data sharing networks. Well positioned organizations such as the GA4GH and the European Data sharing Network could be key players in undertaking efforts here.

[29]Take, for instance, information required by law or governing bodies, to be made available to RECs. In the initial phases of the application and review processes, the basis for deciding what information is relevant is a consequence of established governance, not normative theory. Though those governance practices may have an ethical basis, the mechanisms developed to support them are not, in themselves, normative judgments.

[30]Such as type of consent, whether the participant is an invested party (e.g. whether he or she has a rare disease), or something as simple as the geolocation and jurisdiction of data source.

ing to supply data (Kirkby et al. 2012). It can be further informed by data sharing regulations and best practices currently in use, by recommendations for codification of quality standards for information production and dissemination, (Oleński 2003) by studies on ethical considerations concerning privacy and confidentiality in Big Data research practices (Rajaretnam 2014), and by legal requirements, both present and foreseeable.

3.3 Implementation Strategies

What ELSI concerns are relevant at any given time depends on the context of the research. The goal is not to enforce a one-size-fits-all solution but – and this the great strength of a metadata approach – to make it possible to tailor policy solutions for particular problems and places as required. Once the metadata are in place, it attends the data as it flows from one project to the next. Access to this metadata allows policymakers to more adeptly respond to the needs of biomedical science in each new context and still uphold basic legal and ethical commitments without unduly obstructing research. Below, focussing on consent, I present potential models from which to develop implementation strategies, and discuss some of the benefits of having metadata available for large-scale, international projects.

3.3.1 Learning from Existing Models

To identify which methods are useful for which purposes, it would be instructive to first examine how metadata needs are already being met in areas that utilize biomedical data at the project-specific and the global levels. Project-specific models help to identify in high resolution how data for particular types of research are produced and shared, and what the potential for metadata is. For instance, Genomic Metadata for Infectious Agents (GeMInA), a geospatial surveillance pathogen database, have done work to standardize and integrate heterogeneous sources of information to determine the "who, what, where, when, and how" information surrounding pathogens. Another example is the CancerGrid project, a metadata approach for clinical data management in translational genomics studies in breast cancer. CancerGrid has developed ways to identify appropriate metadata elements for individual datasets that make it possible to annotate, integrate, and query heterogeneous clinical information (Papatheodorou et al. 2009). Methods demonstrated in these examples could prove informative for the "who, what, where, when, and how" of acquisition and use of biomedical data, both in terms of identifying data elements that are useful for ELSI objectives, and in terms of making diverse sources interoperable (Schriml et al. 2010).

Global models help to identify metadata methods that increase interoperability across research contexts. Metadata procedures for coordinating global efforts in genomics have been underway for almost 20 years. Well established web interfaces

already exist which provide a means to enter biological information into these global systems. Once data elements and basic vocabularies have been established through consent matrices or other means, basic ELSI metadata could simply be incorporated into these pre-existing systems, augmenting the biological metadata with ELSI metadata. For example, metadata on basic consent types could be made a requirement of the Joint Genome Institute's Genomes Online Database (GOLD).[31] Initially, the added fields could be basic, such as type of consent attained, if any. Over time, as nomenclatures for describing ELSI factors become more precise, efforts could be scaled up to provide greater granularity. By creating increasingly elaborate consent matrices,[32] these fields could act as points around which detailed conditions surrounding research could aggregate. It would also help to identify where problem areas lie, for instance, when data comes from areas where consent requirements are unavailable or somehow incommensurate across regions and projects.

Similar strategies could be developed for large-scale data sharing efforts, such as the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP),[33] and for initiatives still in their developmental stages, such as Oxford's Big Data Institute,[34] and the GA4GH. Many of these institutions already have robust efforts in metadata development underway to make their data interoperable as they map, coordinate, or consolidate datasets on a global scale. These data environments are considerably more complex than that of GOLD. Metadata procedure developed here may reveal methods for incorporating, aggregating, and harmonising ELSI metadata under similarly complex conditions (Barrett 2012; Inigo et al. 2010).

In some cases, work already done by these institutions identifies what forms of information are most useful. For example, in order to facilitate data sharing among its global federation, the GA4GH provide "consent tools"[35] to ensure basic consent requirements are present which conform to the GA4GH's *Framework for Responsible Sharing of Genomic and Health-Related Data*. Consent tools include advice on "data collected using older 'legacy' consents," language for "where researchers wish to add clauses on international data sharing to actual consents," and "a generic template for new, prospective studies." Together these identify factors which help to support ethical use of data. They ask researchers to confirm whether original consent was obtained and under what conditions. They also ask whether certain factors were foreseen at time of consent, including international data sharing, access

[31] See: https://gold.jgi-psf.org/. At the time this was written, the database hosts information for about 22,000 studies, 67,000 Biosamples, 67,000 sequencing projects and 54,000 analysis projects. "More than just a catalog of worldwide genome projects, GOLD is a manually curated, quality-controlled metadata warehouse" (Reddy et al. 2014).

[32] For a report on an online consent matrix see Thiel et al. (2015).

[33] See: http://www.ncbi.nlm.nih.gov/gap

[34] See: http://www.ndmrb.ox.ac.uk/the-li-ka-shing-centre

[35] See: http://www.p3g.org/news/consent-tools-prepared-global-alliance-genomics-and-health-p3g-ipac

to medical records, genetic/genomic research, or potential commercialization. This information helps researchers to know when ethics committees must be consulted before a person's data is used, and when waivers for re-contact and re-consent may be an option.

These consent tools distil the complicated ethical and legal landscape of international data sharing down to a finite list for a research ethics committee to evaluate: re-contact permissions; confidentiality conditions; security conditions; approved data sharing partners; methods of transference of data; commercial and government uses of data; identification as a member of a population group; when different types of personal data might be combined (e.g. personal data from questionnaires, health data, and genomic data); whether further produced data may be shared (e.g. sequencing of DNA from tissue samples); and conditions under which withdrawal of data from research is possible. Few steps are needed from this list to creating data elements for ELSI metadata. These factors could readily be developed into a consent matrix whose information would assist in identifying conditions surrounding consent in international contexts. Over time, this matrix can come to include more specific information. The specificity requirements could depend upon the region from which the data are sourced, based upon what information gathering requirements are both realistic and appropriate for the region and given the type of research. In absence of metadata like these, it is difficult to identify where, or if, abuses occur. But their presence would allow ethics committees to monitor abuses in data usage and to better identify possible ethical, legal, and social ramifications.

3.3.2 Dataflow

Models like the examples above are useful because they allow flow of metadata on consent to directly reflect dataflow in biomedical research. They help to identify ways to define and integrate the metadata into the dataflow itself. Once clear metadata elements and strategies are identified, checklists can be developed with the aim of establishing community-wide priorities for metadata standards to be used by data producers, research ethics committees, or other governing and oversight bodies. Checklists with standardised languages can be adapted for ELSI objectives, tested, and widely distributed to act as a common denominator across contexts that capture important parameters for policy development (Kolker et al. 2014). A checklist that includes types of consent, for instance, could act as a focal point for increasing granularity of information that would shed light on still more ethical, legal, and social factors research ethics committees and others might want to consider.

Once bases for standardization are established through checklists, consent matrices and ELSI metadata procedures could become a part best practices recommendations for any biomedical endeavour that produces or shares data. Further, the more ELSI related metadata practices can be integrated into the data production of a given institution, the more effective third-party oversight bodies can be in evaluating the ELSI conditions surrounding the data produced by them. Such standards of evaluation would make expectations clearer for researchers and information

providers. They could even create incentives for data producers to proactively self apply high ELSI metadata standards to increase the value and quality of their data for use and reuse.[36]

What information can be gathered from the metadata will depend on many factors, including who is accessing it and for what purpose. Metadata can be made accessible even if the data it describes is not (ECAd et al. 2010). This selective access can allow ethical, legal, and social studies to proceed and evaluations be done without creating additional security risks for the data themselves.[37] Access to this information can assist with developing e-governance systems for managing publicly funded or sourced databanks.[38] It can assist with "ELSI by design" (Kaye 2011) by creating data selection processes that match up suitable data sources with particular research designs. In statistical form, it can be used to track wide scale changes and trends in data as data-enabled science develops (Simeoni et al. 2008).

While the metadata itself may be nonnormative, its analysis could reflect normative standards. It could, for instance, include rules of analysis that match consent type with the appropriate level of risk entailed by the research project. This would be another step in the direction of "ELSI by design." Such rules could help to guide decisions made by legal "safe harbours" (Dove et al. 2013) whose task it is to protect the interests and values of data providers, even when it is being used in regions outside the providers' jurisdictional and cultural contexts. The more granular the ELSI related metadata become, the more they can also assist with establishing and evaluating equivalency principles which help to ensure the normative requirements of one's culture are not violated when data are used in regions which express different values, or express values differently. Requirements of consent, proof of the preservation of individual autonomy, are a prime example of a principle that does not necessarily translate easily to non Western cultures. Here, metadata on type of consent and condition surrounding consent would assist oversight bodies in assuring the ethical, legal, and social functions performed by the principle of consent are still adequately preserved, even when research is done in

[36] A goal of the GA4GH, through a different strategy.

[37] This is not the same as the issue of anonymity of data, deidentification of data, or other methods used to mitigate risks associated with identification. Whether such precautions are effective in Big Data contexts is a matter of debate. This is more closely related to the issue of data security.

[38] Here, examples for managing metadata from areas whose objectives are in some way analogous to those of the ELSI community help. Metadata frameworks developed in e-governance, for instance, have been established to make information accessible to policymakers where and when it is needed. (Inigo et al. 2010**;** Linzer 1988) Metamodeling methods have been developed to increase coordination of data use across government agencies. (Shukair et al. 2013) Additionally, where governance of sensitive data flowing through international contexts is a primary concern, as it is in biomedicine, there may be no better model than finance. When reporting earnings from multiple countries for tax purposes, each jurisdiction has its own regulations. There is high risk of inadvertent non-compliance with local laws. The financial sector has developed intricate metamodelling procedures to address this. They use metadata from multiple jurisdictions at once to create international systems of finance which leave regulations intact at the national level. (Inmon et al. 2008).

cultures that do not share the same individualist perspective. Conversely, when data originate in countries that do not have similar consent requirements, metadata on ELSI factors other than consent can help oversight bodies to determine if those data were collected in ways suitable for countries where consent is ethically and legally required.

Safe harbours and equivalency principles are legal instruments employed to address some of the most challenging aspects of data sharing on the global scale as data sharing and archiving methods integrate project specific data with more universal, interdependent, global visions for biological research. Ensuring information necessary for ELSI concerns attends this data, over the long-term, would allow policymakers to integrate oversight mechanisms at many scales, throughout the many levels of governance, from local research ethics committees to global institutions.

4 Conclusion

As ethical oversight transitions from the needs of the twentieth century to the needs of the 21st, policymakers must identify potential harms as well as potential remedies in ICT contexts. Metadata makes this possible. Those best suited to identify relevant metadata for these complex contexts are those with grounding in ethics, law, sociology, and the history of the changing conditions of governance. They must join with scientists and ICT specialist to examine the complex scenarios and heterogeneous sources of data available, and isolate the information that best informs and directs policy.

Determining what kinds of metadata are helpful begins with understanding the purposes of traditional oversight mechanisms and the histories of their inception, then examining the ways dataflow challenges these foundations. The principle of consent provides a penetrating focal point for doing this. A consent matrix is an efficient method for collecting metadata on conditions surrounding consent. If integrated into best practices now governing consent forms, it would have a harmonizing effect throughout the biomedical community, just as the requirements for consent forms has had.

Access to the metadata this matrix provides would allow research ethics committees, ELSI researchers, and policymakers to better evaluate the conditions around which biomedical data are collected, aggregated, shared, and utilised. It would provide a common platform upon which diverse, often disjointed, policies and procedures developed for the idiosyncrasies of particular institutions and jurisdictions can be better harmonized. This greater functionality would encourage ethical usage of data without unduly impeding access.

The greater granularity of information made available by metadata allows a specificity of oversight and responsiveness that broader, top down, umbrella policies would not likely be able to duplicate. The greater harmonization would help pave the way for better solutions to be adopted where possible, such as dynamic consent.

It would also empower national bodies, such as the National Research Ethics Service,[39] to initiate multicentre research ethics committees that specialise in Big Data biomedical research. These oversight bodies could be supplied with training and tools necessary to interpret metadata in ways responsive to consent conditions. This would preserve the reciprocal relationship that now exists between consenting participants and research ethics oversight bodies, and thereby maintain the trust that is said to make the whole process work.

Acknowledgements The author thanks Dr. Brent Mittelstadt of the Oxford Internet Institute (OII) for stimulating discussion which led to additions and improvements to this chapter, and Dr. Harriet Teare of the Centre for Health, Law and Emerging Technologies at Oxford (HeLEX) for her helpful comments and corrections. The author also thanks HeLEX and Harris Manchester College, Oxford for making this research possible.

References

Altshuler D, et al. 2013. Creating a global alliance to enable responsible sharing of genomic and clinical data. Whitepaper. http://www.google.com/url?sa=t&rct=j&q=global%20alliance%20white%20paper&source=web&cd=1&ved=0CCoQFjAA&url=https%3A%2F%2Fwww.broadinstitute.org%2Ffiles%2Fnews%2Fpdfs%2FGAWhitePaperJune3.pdf&ei=vyVyUrb8NueQ7Aa144DABg&usg=AFQjCNHip6KlFYKzAdRBMG3rbVlnKls2Ow&sig2=PQy2BjkjeRcPNJ9n_Ev86w&bvm=bv.55819444,d.ZGU&cad=rja

Barrett, Tanya, Clark Karen, Gevorgyan Robert, Gorelenkov Vyacheslav, Gribov Eugene, Karsch-Mizrachi Ilene, Kimelman Michael, et al. 2012. BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Research* 40(D1): D57–D63.

Beauchamp, Tom L., and James F. Childress. 2001. *Principles of biomedical ethics*. New York: Oxford University Press.

Carlson, Robert V., Kenneth M. Boyd, and David J. Webb. 2004. The revision of the declaration of Helsinki: Past, present and future. *British Journal of Clinical Pharmacology* 57(6): 695–713.

Caulfield, Timothy, and Jane Kaye. 2009. Broad consent in biobanking: Reflections on seemingly insurmountable dilemmas. *Medical Law International* 10(2): 85–100.

de Vries, Jantina, Thomas N. Williams, Bojang Kalifa, Dominic P. Kwiatkowski, Fitzpatrick Raymond, and Parker Michael. 2014. Knowing who to trust: Exploring the role of 'ethical metadata' in mediating risk of harm in collaborative genomics research in Africa. *BMC Medical Ethics* 15(1): 62.

Dimitropoulos, Linda. 2013. Privacy challenges in health information exchange. *Information privacy in the evolving healthcare environment*, 71. Chicago: Healthcare Information and Management Systems Society (HIMSS).

Dove, Edward S., Bartha M. Knoppers, and Ma'N H. Zawati. 2013. An ethics safe harbor for international genomics research. *Genome Medicine* 5: 99.

Dumbill, Edd, Elizabeth D. Liddy, Jeffrey Stanton, Kate Mueller, and Shelly Farnham. 2013. Educating the next generation of data scientists. *Big Data* 1(1): 21–27.

Dyke, Stephanie O.M., Anthony A. Philippakis, Jordi Rambla De Argila, Dina N. Paltoo, Erin S. Luetkemeier, Bartha M. Knoppers, Anthony J. Brookes, et al. 2016. Consent codes: upholding standard data use conditions. *PLoS Genetics* 12(1): e1005772. http://doi.org/10.1371/journal.pgen.1005772

[39] See: http://www.hra.nhs.uk/about-the-hra/our-committees/nres/

ECAd, Carvalho, A.P. Batilana, J. Simkins, H. Martins, and J. Shah. 2010. Application description and policy model in collaborative environment for sharing of information on epidemiological and clinical research data sets. *PloS One* 5(2): e9314.

Fléchais, Ivan. Designing secure and usable systems. PhD dissertation, University College London, 2005.

Gartner, Richard. 2013. Parliamentary metadata language: An XML approach to integrated metadata for legislative proceedings. *Journal of Library Metadata* 13(1): 17–35.

Inigo, Gil San, Hutchison Vivian, Frame Mike, and Palanisamy Giri. 2010. Metadata activities in biology. *Journal Library Metadata* 10(2–3): 99–118.

Inmon, William H., Bonnie O'Neil, and Lowell Fryman. 2008. *Business metadata: Capturing enterprise knowledge*. Burlington: Morgan Kaufmann.

Karlsen, Jan Reinert, Solbakk Jan Helge, and Holm Søren. 2011. Ethical endgames: Broad consent for narrow interests; open consent for closed minds. *Cambridge Quarterly of Healthcare Ethics* 20(04): 572–583.

Katina, Michael and Mahy, Peter. 2010. An interview with Mr Peter Mahy of Howells LLP who represented S and Marper at the European Court of Human Rights. 153–166.

Kaye, Jane. 2011. From single biobanks to international networks: Developing e-governance. *Human Genetics* 130(3): 377–382.

Kaye, Jane. 2012. The tension between data sharing and the protection of privacy in genomics research. *Annual Review of Genomics and Human Genetics* 13: 415.

Kaye, Jane, Edgar A. Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. 2014. Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics* 23: 141–146.

Kirkby, Helen Michelle, Calvert Melanie, Draper Heather, Keeley Thomas, and Wilson Sue. 2012. What potential research participants want to know about research: A systematic review. *BMJ Open* 2(3): e000509.

Kolker, Eugene, Ozdemir Vural, Martens Lennart, Hancock William, Anderson Gordon, Anderson Nathaniel, Aynacioglu Sukru, et al. 2014. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics: A Journal of Integrative Biology* 18(1): 10–14.

Li Ka Shing Foundation. Report for the Oxford-Stanford Conference on Big Data: Challenges and opportunities for health, 28–29 Nov 2012.

Linzer, Peter. 1988. Is consent the essence of contract – Replying to four critics. *Annual Survey of American Law* (1988): 213.

Lynch, Holly, and I. Glenn Cohen. 2014. Streamlining review by accepting equivalence. *The American Journal of Bioethics* 14(5): 11–13.

McCartney, Margaret. 2014. Care. data doesn't care enough about consent. *BMJ* 348: g2831.

Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: Ethical issues. 2015. https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CCgQFjABahUKEwi36LSw94DHAhUBchQKHcEZBxs&url=http%3A%2F%2Fnuffieldbioethics.org%2Fwp-content%2Fuploads%2FBiological_and_health_data_web.pdf&ei=Zha5VbedHYHkUcGznNgB&usg=AFQjCNHJFMUoBs7giKKxkFuD2D_ofxxzgA&bvm=bv.99028883,d.d24&cad=rja

Oleński, Józef. 2003. The citizens' right to information and the duties of a democratic state in modern IT environment in the light of the UN fundamental principles of official statistics and the ISI declaration on statistical ethics. *International Statistical Review* 71(1): 33–48.

Papatheodorou, Irene, Charles Crichton, Lorna Morris, Peter Maccallum, Jim Davies, James D. Brenton, and Carlos Caldas. 2009. A metadata approach for clinical data management in translational genomics studies in breast cancer. *BMC Medical Genomics* 2(1): 66.

Qin, Jian, and John D'ignazio. 2010. The central role of metadata in a science data literacy course. *Journal of Library Metadata* 10(2–3): 188–204.

Rajaretnam, Thilla. 2014. Data mining and data matching: Regulatory and ethical considerations relating to privacy and confidentiality in medical data. *Journal of International Commercial Law and Technology* 9: 294.

Reddy, T.B.K., Alex D. Thomas, Stamatis Dimitri, Bertsch Jon, Isbandi Michelle, Jansson Jakob, Mallajosyula Jyothi, Pagani Ioanna, Elizabeth A. Lobos, and Nikos C. Kyrpides. 2014. The Genomes OnLine Database (GOLD) v. 5: A metadata management system based on a four level (meta) genome project classification. *Nucleic Acids Research* 43: D1099–D1106. gku950.

Schriml, Lynn M., Arze Cesar, Nadendla Suvarna, Ganapathy Anu, Felix Victor, Mahurkar Anup, Phillippy Katherine, et al. 2010. GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. *Nucleic Acids Research* 38(Suppl 1): D754–D764.

Sheehan, Mark. 2011a. Broad consent is informed consent. *BMJ* 343: d6900.

Sheehan, Mark. 2011b. Can broad consent be informed consent? *Public Health Ethics* 4(3): 226–235. phr020.

Shukair, Gofran, Nikolaos Loutas, Vassilios Peristeras, and Sebastian Sklarß. 2013. Towards semantically interoperable metadata repositories: The asset description metadata schema. *Computers in Industry* 64(1): 10–18.

Simeoni, Fabio, Murat Yakici, Steve Neely, and Fabio Crestani. 2008. Metadata harvesting for content-based distributed information retrieval. *Journal of the American Society for information science and technology* 59(1): 12–24.

Spigner, Clarence. 2007. Medical apartheid: The dark history of medical experimentation on black americans from colonial times to the present. *Journal of the National Medical Association* 99(9): 1074.

Steinsbekk, Kristin Solum, Myskja Bjorn Kåre, and Solberg Berge. 2013. Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? *European Journal of Human Genetics* 21(9): 897–902.

Tan, Jacinta, and Martin Elphick. 2002. Competency and use of the Mental Health Act – A matrix to aid decision-making. *The Psychiatrist* 26(3): 104–106.

Thiel, Daniel B., Platt Jodyn, Platt Tevah, Susan B. King, Fisher Nicole, Shelton Robert, and Kardia Sharon LR. 2015. Testing an online, dynamic consent portal for large population biobank research. *Public Health Genomics* 18(1): 26–39.

Veerus, Piret, Lexchin Joel, and Hemminki Elina. 2013. Legislative regulation and ethical governance of medical research in different European Union countries. *Journal of Medical Ethics* 40(6): 409–413. medethics-2012.

Wallace, Susan E., and Bartha M. Knoppers. 2011. Harmonized consent in international research consortia: An impossible dream? *Life Sciences Society and Policy* 7(1): 35.

Wellcome Trust. 2015. Protecting health and scientific research in the Data Protection Regulation. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf/. A joint statement from non-commercial research organisations and academics. Updated in May 2015.

Willis, Craig, Jane Greenberg, and Hollie White. 2012. Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology* 63(8): 1505–1520.

Willison, Donald. 2003. Privacy and the secondary use of data for health research: Experience in Canada and suggested directions forward. *Journal of Health Services Research & Policy* 8(suppl 1): 17–23.

On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project

Markus Christen, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, and Henrik Walter

Abstract Big Data research is usually explorative, meaning that not all possible hypotheses are known that one may wish to test when data is made available. For the case of biomedical data this poses a significant challenge, as the originators of the data – patients or research participants – have to provide informed consent for using their data. The typically obtained "closed" or "narrow consent", i.e. consenting to use the data in a well-defined research project, is conceptually incompatible with the explorative nature of Big Data driven research. Therefore, "open" or "broad consent" is proposed as an alternative. Nevertheless, open consent cannot justify any type of data use, but requires an "information framework" that separates legitimate from illegitimate Big Data research. For example, consent is given associated with established disease categories: a patient diagnosed with early-onset Alzheimer's disease may consent to his personal medical information being used for any research enhancing our understanding of this particular disease. In our contribution, we address the question whether and how Big Data driven research may undermine

M. Christen (✉)
University Research Priority Program Ethics, University of Zurich, Zollikerstrasse 117, 8008 Zurich, Switzerland
e-mail: christen@ethik.uzh.ch

J. Domingo-Ferrer
Department of Computer Engineering and Mathematics UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain
e-mail: josep.domingo@urv.cat

B. Draganski
Laboratoire de Recherche en Neuroimagerie, Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois, Mont Paisible 16, CH-1011 Lausanne, Switzerland
e-mail: bogdan.draganski@gmail.com

T. Spranger
Faculty of Law/Institute of Science and Ethics, University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany
e-mail: spranger@iwe.uni-bonn.de

H. Walter
Division of Mind and Brain Research, Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Mitte, Charitéplatz 1, D-10117 Berlin, Germany

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_9

this "information framework" of informed consent using the example of the Human Brain Project (HBP). Within the HBP, a Big Data infrastructure is currently being developed to access a multitude of clinical data related to brain diseases based on the conviction that many neurological and psychiatric disorders and diseases are ill-defined in terms of underlying mechanisms. We analyse the interrelation between effects of Big Data research and informed consent and we evaluate ethical and practical consequences.

1 Introduction

Modern biomedical research as well as the ongoing digitalization of healthcare systems is creating an enormous amount of data that has the potential to significantly change our understanding of various diseases. Previous examples of scientific milestones achieved through advances in information technology include the steadily growing number of Internet accessible sequence databases in molecular biology since the early 1980s with its emanation – The Human Genome Project. Neuroscience[1] has clearly taken a similar direction, which is illustrated by several new initiatives for data sharing and common databases. Such initiatives are deemed to be necessary given the massive output of this field. It is estimated that more than 100,000 papers a year are published in neuroscience (Grillner 2014) – most of them involving the analysis of data of various kinds, from genetic data and electrophysiology measurements up to imaging and behavioural data. Compared to other fields like molecular genetics, however, the large majority of neuroscience data sets are still small due to the complexity of the research needed for generating them.[2] Furthermore, data-sharing standards are often lacking. Such small data sets have been referred to as "long-tail" data and may in the future become an important source of new findings (Ferguson et al. 2014).

This traditional focus of neuroscience on "small science" and "small data" comes increasingly under pressure due to recent "big neuroscience" initiatives (Christen et al. 2016). Several Big Data projects are underway to access both small and big data sets generated through research in neuroscience – a development that is exemplified by the "Big Data" issue of *Nature Neuroscience* in November 2014. While many of these efforts focus on model animals, Big Data is also being generated from humans. For example, the amount of openly available and shared neuroimaging data has increased substantially in the last few years (Poldrack and Gorgolewski 2014; Thompson et al. 2014). Even larger data sets concern

[1] In the following, we use a wide understanding of neuroscience, including also medical fields that deal with neurological or brain diseases like neurology, neuropsychology or psychiatry.

[2] Examples include morphological reconstructions of neurons (which is very time-consuming), research with nonhuman primates (which is highly regulated and expensive) or neuroimaging research (which requires a costly infrastructure).

whole-genome sequencing data and the increasing use of technologies for creating large transcriptomic and epigenetic data sets from brain tissue (Shin et al. 2014).

In the following, we will focus on particular Big Data initiatives that are integrated in the Human Brain Project (HBP). The HBP was announced in January 2013 as one of two flagship projects funded by the European Commission's Future and Emerging Technologies Programme. The matched funding for the HBP of about 1.16 billion Euros over 10 years provided by European Union (EU) and partners shall enable a concerted effort to "lay the technical foundations for a new model of ICT (information and communication technologies) based brain research, driving integration between data and knowledge from different disciplines, and catalysing a community effort to achieve a new understanding of the brain, new treatments for brain disease and new brain-like computing technologies" (HBP Report 2012, 3). A major goal of the project involves data integration, for which the HBP is developing six ICT-based platforms dedicated, respectively, to Neuroinformatics, Brain Simulation, High Performance Computing, Medical Informatics, Neuromorphic Computing, and Neurorobotics (for detailed information: https://www.humanbrainproject.eu/). Those platforms are intended to allow sharing research data of all levels of neuronal integration (related, e.g., to ion channel structures, synapse distributions, neuronal microcircuits, brain connectivity patterns, or functional imaging data), methods and models (e.g., in form of computer programs, respectively code) and accessing databases that contain a multitude of clinical data related to brain diseases – the latter will be provided by the Medical Informatics platform and is described in more detail in Sect. 3.

There are certainly many ethical issues associated to data generation (e.g., animal experimentation) and data sharing in neuroscience (e.g., allocation of scientific credit when publishing results originating from shared data). But our focus here is on the problem of informed consent when the data emerges from human subjects, which researchers are required to obtain by current data protection legislation in European countries. Traditionally,[3] "closed" or "narrow consent" is provided, i.e. patients or research participants consent to only one or a few specific uses of the data in a well-defined research project. This, however, is conceptually incompatible with the nature of Big Data driven research that seeks patterns in data based on hypotheses that are often not known when the data has been collected. Therefore, a growing number of researchers and legislators propose "open" or "broad consent" as an alternative, meaning that consent is given to using data for broader research fields or – as a maximum – for any form of research (for an example in genetics, see Lunshof et al. 2008). As we will outline below, such a broad consent poses ethical challenges. These are increased in the case of human brain data, as such data

[3] Seen from a broader historic perspective, (closed) informed consent is a rather recent phenomenon, but can now be considered as standard at least in research settings in industrialised countries. In this contribution, we refrain from outlining the history of informed consent and of international differences in the understanding of informed consent.

is by nature sensitive even if it does not contain healthcare information, because it contains information about the organ of the mind and thus to a certain extent also about the mind itself.

In our contribution we are particularly interested in a potential conflict that is posed by Big Data research in neuroscience, especially when the research is related to neurological or psychiatric diseases. On the one hand, despite the consent being "open", it requires specifying *some information* about what the person is consenting to; otherwise the consent cannot be called "informed". Thus, any form of *informed* consent is embedded in an "information framework" that outlines the general context in which the data is generated, what kind of data is actually obtained, and – although not exhaustively and still in rather general terms – what kind of results could be expected through analysing the data. A plausible and frequent way of generating this information framework is by referring to disease categories – we call this the disease space ontology. For example, a patient diagnosed with early-onset Alzheimer's disease may consent to his personal medical information – health record data, genetic data, neuroimaging data etc. – being used for any type of research enhancing our understanding of Alzheimer's disease.

On the other hand, there is as long-standing discussion in neurology and psychiatry that many current neurological and psychiatric disorders and diseases are ill-defined in terms of underlying mechanisms (Owen 2014; Thagard 2008). On the example of Major Depressive Disorder representing a separate disorder category according to DSM-5, there may be a different classification with a number of subtypes depending on a variety of underlying biological mechanisms. Some types of depressive syndromes may in fact turn out to be other disorders, whereas some might turn out to be subsumed under a disease category with known causes and mechanism and not just a syndrome, i.e. a heterogeneous cluster of symptoms (Monroe and Anderson 2015). Taking these two developments together, it could be that the standard way of providing an "information framework" through disease categories is likely to be shattered through research that necessarily relies on Big Data approaches, in particular in case of brain diseases. We take this apparent paradox as a starting point to explore the connection between the information framework of informed consent and Big Data research that may affect this framework.

This question will be approached in our contribution from various angles. First, we briefly outline the problem of neuroscience-informed disease categorisation with a particular focus on psychiatric diseases. This should motivate the claim that changing the disease space ontology could have an effect on the practice of giving informed consent. Second, we describe in detail the current setup of data collection and informed consent practice within the HBP intended to improve and change our understanding of disease categories in neuroscience. In this way we want to outline that significant changes with respect to our understanding of brain diseases are not a mere theoretical scenario. Third, we discuss the legal problems of open informed consent practices and their dependence on an information framework. In this context, specific attention is paid not only to existing data protection law, but also to legislation aiming at the protection of research participants (e.g. the Council of Europe´s Convention on Human Rights and Biomedicine). Fourth, we evaluate the

underlying moral justifications for upholding or transgressing certain "information borders" in terms of information spheres following the proposals of Nissenbaum (2004) and van den Hoven (2008). Finally, we sketch novel technological solutions for addressing this problem by referring to concepts like traceability of data use and verifiable anonymisation.

2 Disease Categorisation in Psychiatry from a Neuroscientific Point of View

The human brain is among the most complex structures that are object of scientific investigation and it is therefore not surprising that brain diseases are hard to understand. Broadly construed, neurological and psychiatric diseases can be defined as disorders of the brain. There is a continuum of disorders with respect to the degree of their scientific understanding. In some cases, the neurobiological cause is simple and known (e.g. a specific genetic aberration on chromosome 4 in Huntington's disease). Other disorders are diagnostically well-defined and there is a considerable body of knowledge available regarding their underlying mechanisms (e.g. neurodegeneration of dopaminergic neurons in Parkinson's disease). Yet other disorders are difficult to diagnose (in particular in the early phase) and competing theories are available regarding the pathophysiological mechanisms (e.g., Alzheimer's disease). Finally, in many frequent disorders although neurobiological knowledge is available, but rather limited and their definitions today still rely on clinical signs, symptoms and duration (most psychiatric disorders like schizophrenia or depression). In the following, we will focus on the relation between disease categorisation and Big Data driven research for psychiatric disorders, as strong hopes, even promises, have been raised that those approaches can improve knowledge and subsequently therapy (Owen 2014; Wang and Krystal 2008).

According to the two most influential manuals for categorising psychiatric disorders, the IDC-10 of the World Health Organization and the DSM-5 of the American Psychiatric Association, the diagnosis of a psychiatric disorder rest only on clinical features, i.e. on the presence of a specified number of certain symptoms for a specified duration and the exclusion of certain specified causes, like a "organic" disease or an intoxication. The disorder concept of DSM and ICD is categorical: either you have the disease or you don't – although disorder are characterised by different degrees of severity, e.g. for depression. ICD-10 as well as DSM-5 do not rely on underlying pathophysiological mechanisms as most of them are not known, heavily debated, or can only be diagnosed post-mortem.

Between the publication of DSM IV (released in 2004) and DSM-5 (released in May 2013) it was hoped that the new DSM-5 would advance the field considerably with respect to two issues: integrating dimensional approaches (i.e. use constellation of symptom dimensions instead of categories for example for the diagnosis of personality disorders) and integrating neurobiological criteria (genetic, molecular, neuroimaging) for making diagnoses. Suggestions in this direction were intensively

discussed by the DSM-task force of the American Psychiatric Association over several years. At the end, however, none of these conceptual changes were included in DSM-5. This was largely because it was felt that neurobiological knowledge was not (yet) reliable enough, but also due to the fact that the DSM is much more than a medical nosology: it also serves a central societal role by providing the basis for mental health care and thus is conservative in nature as changes would immediately affect millions of patients and carefully balanced systems of providers and consumers.

This missing integration of neurobiological knowledge frustrated many mental health scientists. In fact, 3 weeks before the official release of DSM-5, Thomas Insel, at that time director of the National Institute of Mental Health (NIMHM),[4] the largest research institute for mental health in the western world, launched a considerable attack on DSM-5 by declaring in his blog that "the weakness (of DSM-5) is its lack of validity. Unlike our definitions of ischemic heart disease, lymphoma, or AIDS, the DSM diagnoses are based on a consensus about clusters of clinical symptoms, not any objective laboratory measure. In the rest of medicine, this would be equivalent to creating a diagnostic system based on the nature of chest pain or the quality of fever. Indeed, symptom-based diagnosis, once common in other areas of medicine, has been largely replaced in the past half century as we have understood that symptoms alone rarely indicate the best choice of treatment" (Insel 2013). The apparent lack of availability of reliable biomarkers for mental disorders was explained by Insel as a conceptual rather than as an empirical problem: it would be equivalent to rejecting the usefulness of the electrocardiogram (ECG) as a diagnostic tool, only because many patients with chest pain do not have ECG changes. In fact it was the ECG which allowed differentiating chest pain due to specific heart problems from other forms of chest pain, i.e. the tool helps to categorise the disorders by measuring physiological processes. And, according to Insel, the same should be done in psychiatry by "collecting genetic, imaging, physiological, and cognitive data to see how all the data – not just the symptoms – cluster and how these clusters are related to treatment response" (Insel 2013). Such an approach is only possible using Big Data techniques, as we will outline in the next section.

In fact, the NIMH started a research program some years ago which is now known under the name Research Domain Criteria (R-DOC; Morris and Cuthbert 2012). The basic idea of this approach is to achieve a dimensional characterisation of mental illness as mentioned above in order to discover, refine or reclassify mental disorders. For this purpose, it is suggested to study diseases based on a two-dimensional grid based on current neurocognitive and molecular approaches and knowledge. One dimension consists of five core domains of mental functioning ("systems") that have been determined by consensus conferences of active scientists

[4]Interestingly, in particular with respect to the increasing role of ICT for (mental) health, Thomas Insel announced in September 2015 after 13 years serving as director of the NIMH that from November 2015 on he will move to Alphabet, the umbrella organization of Google in order to help to develop mobile health technologies.

from the field, i.e. systems for negative valence, positive valence, cognition, social processes and arousal. Each of these domains has subdomains, e.g. the system for negative valence comprises the subdomains active threat ("fear"), potential threat ("anxiety"), sustained threat, loss and frustrative non-reward. The other dimensions refer to levels of organisation on which the constructs within the domains can be measured: from genes, molecules, cells, circuits, physiology, to behaviour, self-reports, and paradigms. By filling this 2-dimensional grid with scientific results, it will, in the long run, be possible to characterise mental disorders on a sound empirical basis and detect patterns leading to the discovery of new disorders or reclassification of new ones. These discussions on a new understanding of mental disorders as disorders of neurocognitive domains are also referred to as the "third wave of biological psychiatry" (Walter 2013).

However, for this approach to being realised, a revolution, or at least a reform of disease concepts is required. It also would entail Big Data neuroscience on mental health: only if you have obtained enough high dimensional data from many domains of many subjects together with clinical data, this approach might become successful. But standard DSM-based research uses the (not-so) gold(en) standard of symptom-based categories and will thus make no progress. Therefore, Insel has announced that the NIMH will in the future not fund research based on "old" still gold-standard disease categories, but rather RDOC-oriented, dimensional research. To take a simple example, it would not fund neurobiological research on alcohol addiction, but rather neurobiological research on impulsivity as a contributor to alcohol drinking.

But what would such a change induce on the level of actual researchers who have to interact with patients and research subjects and obtain their informed consent for using their data? Consenting to the use of data in research obviously requires a basic understanding on the context in which the data has been generated and in which it is likely to be used. Lay people like patients usually do not have the competences needed to assess the detailed hypotheses of research in which they are involved, e.g. when they are asked to participate in a clinical trial for testing a new medication. Although such detailed information is not required, as the main interest of the patient probably is to obtain information on possible health risks and benefits – this type of information is still presented in a context framed by the disease from which the patient is suffering. Taking the simple example from above, a patient with a severe drinking problem would probably expect that the research in which he is involved relates to alcohol addiction and not to some research on impulsivity, as the person may consider impulsivity (to some degree at least) as a legitimate aspect of his personality. Thus, the specific disease along with a laymen understanding of what, e.g., a depression or Alzheimer's dementia involves, is crucial for putting the informed consent into a context.

This context also affects the moral significance of diseases. A disorder caused by a genetic factor (e.g. Huntington's disease) is associated with specific types of moral problems (e.g., related to inheriting the disease) that are not perceived to be present in neurodegenerative disorder. Some brain disorders are associated with a stronger stigma than others (e.g. schizophrenia versus epilepsy). Yet some disorders are

understood to be clearly "brain based" (e.g., Parkinson's disease), whereas others are much more associated to "external" (e.g., social or cultural) causes, although it is likely that changes in the brain play an important role in the disease course (e.g., anorexia nervosa) – and such "external causes" involve a different responsibility relation (e.g. by avoiding certain social settings or by generating an imperative to change certain societal aspects through policy interventions). Consenting to use data related to one disease may thus mean something different than consenting to contribute data related to another disease or to broader spectrum of diseases relevant to specific domains of functioning.

If now a research program is installed that seeks connections between neuronal diseases that lay people consider rather different, should they be informed on these possible links? For example: should a Parkinson's disease patient be informed that analysing her data may help to understand schizophrenia or depression – and in this way implicitly given her some reason to suspect that she might suffer also from one those diseases? Actually, the re-conceptualised disease space ontology may look very different compared to the disease space that frames the current social handling of these diseases in terms of physician-patient relation, health insurance, or stigmatisation. Here, we try to sketch possible ethical consequences of such a change in the disease space. But before that, we outline the actual possibility that such a change could happen (Sect. 3) and the current legal setting related to informed consent (Sect. 4).

3 Data Collection, Informed Consent and the Human Brain Project

Every day, an impressive amount of data related to brain health and disease are produced in clinical and research establishments across Europe. Usually, these data are in the format of descriptive clinical data, laboratory results or brain images that serve to help medical decision-making. They are viewed mostly only once before being archived on departmental or laboratory servers for a finite number of years. This mass of data constitutes an enormous research resource that is currently largely unused. Though the data are collected at different sites, it has now been demonstrated that the variance introduced by analysing data from multiple imaging platforms or clinical chemistry laboratories is much smaller than the variance that is attributable to the disease (Stonnington et al. 2008). In other words, variability through differences in methodological practice can be controlled. This fact suggests an opportunity to use archived data for the pathophysiological, anatomical and medical studies on a population basis. This is a major motivation of the data integration strategy of the Human Brain Project (HBP) in the medical informatics platform.

Recent advances in computing and commercially available algorithms for federating data from local databases that work unobtrusively in the background in real

time make such a project practicable and cost-effective. The Medical Informatics platform of the HBP proposed an initial programme based on federation of data related to brain diseases to establish feasibility, sharing protocols, data usage agreements, access protocols and other issues. This idea represents a quantum leap from the path trodden out in the past by successful database initiatives such as the *Alzheimer's Disease Neuroimaging Initiative* (ADNI, see www.adni-info.org), which is used by many researchers world-wide although it is much smaller in scope and more expensive because the technology was not available at the time of its setup.

From an ethical point-of-view, the mass of brain health and disease related information collected in hospitals, clinics and research establishments is grossly underused at present, which represents an extraordinary waste of resources. With advances in modern information technology, especially in terms of massive data storage and access to hardware, analysis tools and data mining techniques, these data can be used to carry out a range of studies of social and medical importance. The range of possible investigations is enormous, if the data can be systematised and intelligently mined. The main goal of the Medical Informatics platform of the HBP is to federate and integrate clinical and basic science research together with information technology and establish new ways for open access to shared aggregate data in order to ask hypothesis driven questions, to mine data, to carry out epidemiological, genetic and other surveys. But certainly, the question emerges whether the practice of large-scale access to this data is compatible with the informed consent given by the patients from which this information emerges. We will come back to this point later and we first outline the technical procedures of data collection.

The complexity of clinical data especially in the field of neuroscience makes evident the need for a coherent framework for integrating the multiple temporal and spatial scales of data to facilitate its interpretation. Ongoing large-scale projects (e.g., ENIGMA, Human Connectome Project, Allen Brain Atlas, GENSAT) demonstrate that brain imaging data capturing *in vivo* anatomical and functional information about the brain can serve as a backbone for developing a viable framework for research data integration. From a clinical perspective, the more prominent examples are the recent developments in the neuro-epidemiology of dementia based on differential patterns of cortical atrophy associated with cognitive decline; the development of biomarkers from analysis of scans and subsequent cognitive outcome or neuro-pathological examination; population wide genetic association with *in vivo* pathology studies, as demonstrated by image-derived brain tissue characterisation.

To give a very specific example that illustrates what could become possible, there is a pressing demographic and economic need to answer questions about the preclinical stage of dementia, in particular the incidence and natural history of pathological change, early detection and diagnoses based on brain measures rather than behavioural expression, and how to monitor the rate of pathological brain changes on sufficient numbers of people such that the results are generalizable. The repercussions of the results will be important because there is preliminary evidence

to suggest that the dementias can be differentiated by early distribution of brain atrophy. It should therefore be possible to identify purer cohorts of the different dementia-associated diseases than is now possible to identify, and to test and develop new specific disease-modifying drugs. This type of research could eventually lead to a new classification of dementia-associated diseases that is quite distinct from today's understanding. This would be an example of a re-conceptualized disease space through Big Data research.

The use of data mining – the technological precondition for restructuring the disease space – involves the extraction of patterns from large sets both for scientific and business related queries. The use of this technique has exploded in the last few decades in many fields in biomedicine, as outlined in Sect. 1. Considering the remarkable advances in biomedical imaging technology and analysis, data mining offers new opportunities capitalising on the ability to extract characteristic features from abundant and diverse information about human (patho-) anatomy and physiology. The creation of disease-specific neuroimaging data repositories (ADNI, ENIGMA, IMAGEN) represents first attempts to use advanced neuro-informatics methodologies for databases of clinically relevant information. Although offering standardised data processing of anatomical brain images, these databases serve mainly as repositories rather than frameworks for data mining on clinical neuroscience grounds. Data mining approaches are aimed at making use of the large data set in order to extract main predictors that explain variance in the data. An understanding of the nature and extent of inter-subject variation is critical for the characterisation of the neural basis of cognitive processes in healthy subjects and the changes that cause abnormal functioning. Data mining approaches build upon the decomposition of inter-individual differences to create meaningful classifications of subjects and predictions of continuous variables such as behaviour or performance. The principal hypothesis is that characteristic distributions of variability of the structure of the brain and its connectivity patterns will be of diagnostic value through identification of disease discriminative patterns.

The Medical Informatics platform of the HBP (see www.humanbrainproject.eu/medical-informatics-platform; Frackowiak and Markram 2015) is building on a concept for data federation that allows mining all available resources without the need to directly access the original data. Rather than copying, downloading and mirroring data, the current set-up focuses on locally creating data aggregates, which provide a summary of the available data at a particular site. These aggregates are feature-specific and can be queried by the end-user in the form of double-aggregated data. At no instance is there access to individual-specific data, which could open the possibility for data misuse and identification of a given person. The combination of simple database language queries and advanced methodological tools for statistical inference and learning allows harvesting the aggregated data, binding multiple sources of information and extracting characteristic features to answer domain-specific questions. This is a dynamic process that aims to create clinical generative models of specific diseases. As more data is gathered models can be re-evaluated and refined to answer more subtle questions.

The processing of the data to extract features for aggregation is performed locally within the secured systems of hospitals based on the availability of powerful algorithms able to handle vast amounts of data. Several issues surrounding big data analysis on the Medical Informatics platform of the HBP need consideration in relation to data protection. Legally binding laws enforced by the EU authorities stipulate that responsibility for the data and its ownership is transparent (see also Sect. 4). Although our framework does not allow accessing individual data, current laws and regulation apply to data transfer, data processing and data security and the Human Brain Project has to ensure that data management is compliant with data protection law. This also means that beyond strict procedures for data anonymisation, data preparation for mining should be restricted to well-defined workflows that prevent data miners from identifying specific individuals or from uncovering confidential information. This protection of privacy is – from an ethical point of view – the uncontroversial part of the problem. It is also in the focus of current legislation, as we outline below. But our question is, whether privacy is the only and main concern of Big Data driven research in neuroscience.

4 Legal Issues of Open Consent and Its Information Basis

The prevailing Data Protection Law applicable in all EU Member States is mainly based on European legal guidelines. Debates over the minutiae of a new EU Data Protection Regulation (anticipated to be passed around 2016) are fully underway.[5] Particularly the question of how this new Law will affect the use of personal data in a scientific context is one of the main aspects in need of clarification. However, there is a large consensus that the mere, indiscriminate adoption of general data protection standards for scientific work could pose an unnecessary and unjustified restriction of the freedom of research.

Furthermore, specific problems arise regarding the practical implementation of so-called informed consent. Originally developed as a bioethical principle, the notion of informed consent can be found in many documents of international and national law (e.g. the Council of Europe's Convention on Human Rights and Biomedicine) today and has thus become an integral part of the Positive Law. In 'classical' research projects such as in clinical trials it has been shown, however, that providing (too much) information to the research subject can occasionally lead to the opposite effect of what the informed consent aims at; excess of information can leave the concerned party unable to make a (truly) informed choice after all. This problem exacerbates in Big Data driven research areas as outlined above: one cannot effectively communicate the potentially enormous range of testable hypotheses to patients. Therefore it has to be examined which models of informed consent can

[5] An overview on the legislation procedure is available here: http://ec.europa.eu/justice/data-protection/

be used or further developed, that protect research participants on the basis of legal compliance and yet do not disproportionately restrict the efficiency of research.

However, integrating informed consent into Big Data driven research also touches upon the question whether or not the concept as such leads to an adequate protection of research participants. Especially with regard to other contexts, e.g. establishing so-called bio-banks, it was and still is discussed, if informed consent in its classic understanding is sufficient to protect the rights and interests of research participants in a sufficient manner (D'Abramo 2015; Hofmann 2009).

Due to the complexity and high dynamics of modern biomedical research, it is stressed that research participants may not realise the full implications of giving their consent. While agreeing to the use of their samples or test results for 'the purpose of research', participants may have a lack of understanding what exactly that vaguely phrased expression means (Cordasco 2013). Therefore, in a legal context, the restriction of the range of the informed consent has been consistently demanded. A restriction may be imposed with regard to a certain time frame (e.g. informed consent is given for a time period of 5 or 10 years, in combination with the obligation to newly clarify the purpose of the use of the elevated personal data in order to attain a renewed consent of the participants) or regarding a factual aspect which would restrict the given informed consent to a particular project or to the research of a specific disease pattern.

However, the approaches of restricting either the temporal or the factual context of informed consent fail to work even with regard to small-scale projects: a renewed informed consent can usually not be obtained due to research participants moving away or dying in the meantime. Furthermore, the maintenance of an address register would not only go beyond the scope of time effort, but also be a great financial burden to any project and may actually generate new privacy risks due to problems in securing this information from unauthorized access.

A restriction of the informed consent to a certain factual context is problematic as well, because Big Data research aims for a cooperation and combination of different projects, and not for individual projects. In addition, undertaking a follow-up project would be made impossible for the researcher who got the informed consent in the first place. Lastly, the restriction to only one specific disease pattern is problematic as well due to the difficulty of insufficient clarity and changing definitions and understandings of a certain disease – as we have outlined in detail in Sect. 2.

Regarding the reasons mentioned above, jurisprudence represents a general permissibility of a 'broad consent' which is of unlimited time and enables largely unrestricted factual research. As far as some legal systems assume an inadmissibility of a 'general consent', this concept deals with the consent given by a third person to carry out any kind of legal action and cannot be compared with the approach and content of a 'broad consent' (see for the case of biobanks: Serepkaite et al. 2014). The latter does not mean that contributors of genetic material or data do not obtain any rights. Personal rights and data protection laws as well as privacy issues obviously have to be respected. Therefore, the current legal understanding of the problem of Big Data driven research focuses on demanding technological solutions that ensure that privacy and data protection are respected, mainly through

aggregation and anonymization techniques or – more generally – privacy-by-design approaches. This intermediate conclusion from a legal point of view leaves open two questions: First, are these technologies actually able to protect the privacy of the research participant? And second, what are the deeper moral reasons and possible effects of ethical significance when such a new 'broad consent' regime is implemented? We will now focus on the second question.

5 Ethical Issues of Changing the 'Information Framework'

When assessing the problem of informed consent in a Big Data context, a historical perspective is helpful. The notion of informed consent has been put in the centre of bioethical considerations after one of the darkest episodes in the history of medical research – the horrific experiments carried out by doctors on concentration camp victims in Nazi Germany. In the Nuremberg trials of 1947, the requirement that "The voluntary consent of the human subject [to medical research] is absolutely essential" has been formulated for protecting the participant from harm. These requirements strongly influenced the Declaration of Helsinki, that later underwent several revisions, in particular related to the notion of informed consent (Carlson et al. 2004). Despite these changes, the 'moral core' of informed consent in the bioethical common-sense-understanding is protecting the individual from involuntarily incurred harm. From that perspective, the ethical question is, whether Big Data driven research backed by 'broad consent' could create additional harm for the subject – i.e. harm not directly related to the research intervention itself (e.g., the risks of some imaging techniques, which certainly are part in the information procedure when obtaining informed consent), but to long-term outcomes of the research. As the current legal discussion described in Sect. 4 demonstrates: the focus of the discussion is almost exclusively on privacy breaches as the main harm that could result. For example, one wants to avoid that the genetic data of a person with Huntington's disease made available for research can lead to a re-identification of this person, thereby harming this (still healthy) person in her social setting, e.g., by provoking a dismissal from her job.

We certainly do not dispute that this kind of harm is of relevance in Big Data driven research – and the main ethical question here is whether the technological solutions for preventing such harm actually will do their job. This aspect will be further discussed in Sect. 6. But we suggest two further issues that need ethical consideration: First, the necessity of broad consent for Big Data driven research may pose additional problems that have harm-implications. Second, broad consent is associated with other (positive) ethical values than harm-prevention that may help to make Big Data research more ethical. We will now discuss these two issues in more detail.

The first issue relates to the point that providing informed consent requires informing the patient on the intervention that will generate the data. On the one hand this concerns information on the direct risks and consequences of the intervention

itself. For example, in case of a MRI scan of the brain, issues like technical risks (e.g., implants) or incidental findings have to be discussed. This part of the informed consent procedure is not affected by a subsequent use of the data in a Big Data context. On the other hand, the patient has to be informed at least to some degree on the potential use of the data. If the informed consent is broad, this degree will be quite unspecific, but still needs some framing. Patients are unlikely to accept an explicit formulation like 'You agree that your personal data will be used for any kind of application'. Thus, a framing in two respects will be necessary: First, one has to induce trust in the patient that harm through privacy breaches will effectively be prevented. Second, some factual framing will be needed. Probably the broadest kind of factual framing is that the data will be used for *research* purposes (and, e.g., not sent to a wellness company such that they can tailor new commercial offers for patients with similar diseases). More likely is, however, a (at least implicit) framing that the data will be used for research related to the medical condition of the patient. But why is such a framing necessary?

The reason for this is – as we suggest – that information frameworks play a decisive role for giving moral meaning to the world we live in. This insight can be partly attributed to the idea of *spheres of justice*, introduced in 1983 by the philosopher Michael Walzer, which proposes that societies consist of different social spheres (e.g., medical, political, market, family and educational) each defined by a different type of good that is central to that particular sphere (Walzer 1983). These different types of goods (e.g., medical treatment in the medical sphere, political responsibility and public office in the political sphere) and the meaning and significance they have in each of these spheres, have their own associated criteria, principles and mechanisms concerning their distribution and allocation. In order to prevent mixing up distributional criteria and goods from different spheres (and prevent, e.g., allocating seats in parliament on the basis of financial assets, family relationships or health condition, or making one's ranking on a waiting list in health care dependent on family relationships or college degrees) these spheres have to be kept separated. This idea implies amongst other things that the distribution of access to particular goods tracks the sphere's specific normative considerations (e.g., 'need' in the medical sphere, 'democratic election' in the political sphere). Goods have to be distributed along the mechanisms of the corresponding sphere and goods from different spheres ought not to influence each other in terms of distribution. Put differently, this means that the exchange of goods between spheres has to be "blocked"; Walzer talks about "blocked exchanges" and the "art of separation".

Walzer's work has been applied to the realm of information systems by Nissenbaum (2004) and Van den Hoven (2008). Nissenbaum coined the term *contextual integrity* of social spheres, whereas spheres are defined through the expectations and behaviour of actors that differ per sphere. In order for contextual integrity and sphere separation to be achieved, the type of information that is revealed and the flows between different parties have to be appropriate for the context.

Within the broader privacy debate, the challenge of Big Data is that information produced within these spheres (health, politics, criminal justice, market) travels much faster and is more difficult to control than in the traditional offline world.

So we face a set of phenomena that threaten the integrity of social spheres and the cultural and social meanings expressed in them, including our values. Of course the boundaries between spheres are to a certain extent relative to time and culture, and not carved in stone forever, but it is important to note that every age, society and culture does in fact draw and treat these boundaries – construed as sets of constraints on the flow of information – as of high normative relevance.

Going back to our example, this also means that broad consent should respect these boundaries. The point is that providing broad consent for using data can transgress these boundaries in ways that generate indirect harm for the person who provides the data *even in cases when privacy is fully respected*. For example, researchers emerging from fields completely unrelated to the disease condition of the patient may use the (aggregated and anonymized) data to check for connections between health conditions and credit rating; resulting finally in a policy that prevents the patient in future to obtain certain bank credits. This would be considered a breach of a boundary between two social spheres with quite different moral regimes: the health sphere on the one hand and the economic sphere on the other hand. Other researchers may use the data in a way that finally results in a genetic test that allows testing foetuses – and in this way offer the option of abortion to the future parents. Such a development may be against core-values of the patient when she reads about this type of research in the newspapers, as she realises that her data may have played a role in this research. In this case, personal boundaries between acceptable and unacceptable applications of scientific research are breached. Yet other researchers may – based on research that includes the anonymised data of the patient – come to the conclusion that some sub-form of a neurological disorder (actually the condition from which the patient suffers) is associated with another disease that has a much stronger social stigma – and the patient is finally confronted with social exclusion resulting from the public dissemination of this reconceptualised disease space.

The underlying problem of these still hypothetical cases is that through broad consent, the consenting person risks that his data finally leads to research result that transgress important moral boundaries of this person or of society in general. The person contributes to a "new world" which he personally rejects. Thus, the question emerges how broad consent can be made compatible with respecting these boundaries.

Answering this question involves the insight that requiring consent is not merely an act to protect a person from unwanted harm – the classic understanding of informed consent. But it also involves requiring an explicit agreement to contribute to something that the person considers to be a valuable goal. Consenting is an act of autonomy that has a positive motive (e.g., compassion) and is backed up by some understanding of fairness (e.g. that the resulting research is not leading to unjustified discriminations). Understanding consent as such an active act entails the notion of responsibility in two ways: First, the consenting person trusts that the researcher will deal responsibly with this data – both with respect to preventing privacy breaches as well as with respect to the goal of the study. Second, the consenting person may to some degree be set in a position to control the usage of the data. Although it will probably be an exception that the person herself would like

to track the usage of her data, one may consider a model of "data stewardship", i.e. an institutional setting that (as a representative of the data provider) allows tracking data usage and regularly report on how the personal data of people has contributed to research. Both ensuring trust and responsibility will have to be "materialised" through technological solutions that function and that can be understood by both the users of Big Data technologies as well as those who provide the (Big) data. Whether these technologies are available is the topic of the next section.

6 Technological Ways of Securing Open Consent

A major technological problem related to this aim of enhancing trust and responsibility is that current anonymisation practice does not take the informational self-determination of the data subject into account. Since in most cases the data releaser is held legally responsible for the anonymisation (for example, this happens in official statistics), the releaser favours global anonymisation methods, where he can make all choices (methods, parameters, privacy and utility levels, etc.).

When asked to provide data and consent, the subjects must hope there will be a data protector who will adequately protect their privacy in case of release. Whereas this hope may be reasonable for data collected by the public health care system or more generally by (democratic) administrations, it may be less founded for private surveys (data collected by pharmaceutical companies or by any other private company). Indeed, a lot of privately collected data sets end up in the hands of data brokers (U.S. Federal Trade Commission 2014), who trade with them with little or no anonymisation. Hence, there is a fundamental mismatch between the kind of subject privacy (if any) offered by data releasers/protectors and privacy understood as informational self-determination: usually, the subject is not given control on how her data is protected.

To empower the data subject, Domingo-Ferrer and Muralidhar (2015) proposed a permutation-based paradigm of data anonymisation. They showed that any anonymisation method is functionally equivalent to permutation plus (perhaps) a small amount of noise. In a nutshell, if one compares the ranks of the values of each original and each anonymised attribute, one finds that the effect of any anonymisation method is to change the ranks to some extent, which can be viewed as a permutation (see Domingo-Ferrer and Muralidhar (2015) for more details and a running example). Based on this, they defined a new privacy model, called ($\boldsymbol{d}$, $\boldsymbol{v}$, $\boldsymbol{f}$)-permuted privacy that is verifiable by the subject. When given the anonymised data set by the data protector, each subject can check how much the values in her record have been permuted and whether this permutation is sufficiently protective.

Just allowing the subject to verify protection may not be enough or even worse than not allowing verification if the subject is left unsatisfied with the level of protection provided. An unsatisfied subject may refuse to answer and/or to give consent the next time the data collector approaches her. A more constructive alternative would be to allow the subject to take care of the anonymisation of

her own data record (*local anonymisation*, e.g. Song and Tingjian Ge 2014). In the context of the HBP, in some cases it may be viable for patient subjects to use their personal devices (e.g. smartphones) to conduct local anonymisation. For example, if a patient is being continuously monitored through sensors connected to her smartphone while at home, clearly all data being collected can be locally anonymised by her smartphone.

Beyond assuming a well-informed subject with some basic knowledge of the implications of anonymisation, a problem of local anonymisation is that the subject must anonymise her record without seeing the records of the other subjects. Hence, the subject cannot know whether the anonymisation she is applying will permute the values of her record enough with the values of the other subjects. To play it safe, each subject is likely to add a lot of noise to her values, which results in an anonymised data set with too poor utility.

In Soria-Comas and Domingo-Ferrer (2015) *collaborative anonymisation* has been proposed as a synthesis alternative that seeks to empower the subjects while preserving data set utility as in the case of centralised anonymisation for the same privacy level. The idea is that subjects generate the anonymised data set in a distributed and collaborative manner. Neither the data collector nor subjects gain more knowledge about the confidential information of a specific subject than disclosed by the anonymised data set.

Let us analyse the motivations of a rational subject to engage in collaborative anonymisation. Rationally, she will only contribute to form an anonymised data set if the benefits she obtains compensate her privacy loss:

- A subject without any interest in the research made possible by the data being collected is better off by declining to contribute (privacy prevails). Note, however, that subjects may have indirect interests, like expecting a potential benefit from the research conducted with the data (better healthcare, better life conditions, etc.) or simply satisfying a philanthropic inclination.
- A subject without privacy concerns can directly supply her data without any anonymisation requirements (potential benefit prevails).
- A subject who is interested in the research made possible by the data but has privacy concerns should prefer the collaborative approach to both the centralised and the local approaches because: (i) It outperforms centralised anonymisation by offering privacy with respect to the data collector; (ii) it outperforms local anonymisation because it yields less information (utility) loss and hence enables better research.

Collaborative anonymisation leverages the notion of co-utility (Domingo-Ferrer et al. 2015), which refers to protocols (interactions) designed in such a way that the best strategy for a rational selfish player to attain her goal is to help some other players to attain theirs. Co-utile protocols make mutual help self-enforcing. In anonymisation of individual data, the privacy protection obtained by a subject positively affects the protection that others get. In other words, when masking the identity of a subject within a group, none of the subjects in the group is interested in making any of the other subjects re-identifiable, because that makes her own data

more easily re-identifiable. In this sense, we can say collaborative anonymisation is co-utile. Specifically, Soria-Comas and Domingo-Ferrer (2015) give a co-utile protocol to achieve k-anonymity in a collaborative way. *k*-Anonymity is a privacy model in which each subject in indistinguishable within a group of *k* subjects when looking at the released data set.

While the above *(* $\boldsymbol{d}, \boldsymbol{v}, \boldsymbol{f}$ *)*-permuted privacy model can allow a patient/subject to verify how well her data have been anonymised, and local/collaborative anonymisation can give the subject full control on the anonymisation process, privacy is not all a patient may need, as mentioned in Sect. 5. Being able to track the usage of her data is a complementary (and probably more ambitious) requirement. In fact, for some types of data used in HBP, anonymisation may be unfeasible because the data is inherently identifying and cannot be altered to make it less identifying (e.g. this is the case of genetic data or even human brain scans); for such data, all the patient could be promised by the researchers/collectors is to keep track of who accesses it and how it is used (the data stewardship mentioned in Sect. 5). Such tracking is addressed by the so-called provenance technologies. Provenance refers to the chain of successive custody (including sources and operations) of information (or even hardware equipment). The current practice of information provenance is rather rudimentary and still far from being dependable enough. The good side is that there are many sectors interested in improving provenance technologies: beyond healthcare research and HBP, the banking sector, the software industry and the cybersecurity sector are important fields where tracking information usage is very important. Hence, substantial research efforts are underway: technologies have been demonstrated for annotation in scientific computing, for provenance-aware data storage (automatically tracking accesses, downloads, etc.), for building tamper-resistant chains of custody, for pedigree management (tracking the source of data), etc. (See Chapter 9 of U.S: Homeland Security 2009 and references therein).

7 Conclusion

In conclusion, we suggest that a broader ethical focus would allow understanding the ethics of Big Data driven research not solely as an issue of upholding the privacy of the individual who consents to her data being used, but also as a matter of individuals that decide to contribute to a positive goal and thus would like to be put into a position such that they can trust that they are indeed making the world a better place. This requires generating an understanding on how Big Data research may affect the ontology upon which consent decisions are based (e.g., disease ontologies) as well as the underlying, morally significant boundaries. This also requires developing and integrating technologies in Big Data research that empowers the subject so that she really is in control of what happens to her data before it is released. This should enable her to give her data and her consent in conditions that are more compatible with informational self-determination.

References

Carlson, Robert V., Kenneth M. Boyd, and David J. Webb. 2004. The revision of the declaration of Helsinki: Past, present and future. *British Journal of Clinical Pharmacology* 57(6): 695–713.

Christen, Markus, Nikola Biller-Andorno, Berit Bringedal, Kevin Grimes, Julian Savulescu, and Henrik Walter. 2016. Ethical challenges of simulation-driven big neuroscience. *AJOB Neurosci* 7(1): 5–17.

Cordasco, Kristina M. 2013. Obtaining informed consent from patients: Brief update review. In: Making health care safer II: An updated critical analysis of the evidence for patient safety practices. Rockville: Agency for Healthcare Research and Quality (US) (Evidence Reports/Technology Assessments, No. 211) Chapter 39. Available from: http://www.ncbi.nlm.nih.gov/books/NBK133402/

D'Abramo, Flavio. 2015. Biobank research, informed consent and society. Towards a new alliance? *Journal of Epidemiology and Community Health* 69(11): 1125–1128. doi:10.1136/jech-2014-205215.

Domingo-Ferrer, Josep, and Krishnamurty Muralidhar. 2015. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. Technical report. Available at: http://arxiv.org/abs/1501.04186

Domingo-Ferrer, Josep, Jordi Soria-Comas, and Oana Ciobotaru. 2015. Co-utility: Self-enforcing protocols without coordination mechanisms. Proceedings of the 2015 international conference on Industrial Engineering-IEOM 2015, Dubai, 3–5 Mar, 1–7.

Ferguson, Adam R., Jessica L. Nielson, Melissa H. Cragin, Anita E. Bandrowski, and Maryann E. Martone. 2014. Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience* 17(11): 1442–1447.

Frackowiak, Richard, and Henry Markram. 2015. The future of human cerebral cartography: A novel approach. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. 370(1668) pii: 20140171. doi:10.1098/rstb.2014.0171.

Grillner, Sten. 2014. Megascience efforts and the brain. *Neuron* 82(6): 1209–1211.

HBP Report. 2012. The Human Brain Project. A report to the European Commission. Available at: https://www.humanbrainproject.eu/documents/10180/17648/TheHBPReport_LR.pdf/18e5747e-10af-4bec-9806-d03aead57655. Accessed 4 Sept 2015.

Hofmann, Bjørn. 2009. Broadening consent – And diluting ethics? *Journal of Medical Ethics* 35(2): 125–129.

Insel, Thomas. 2013. Transforming diagnosis. Available at: http://www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml. Accessed 3 Sept 2015.

Lunshof, Jeantine E., Ruth Chadwick, Daniel B. Vorhaus, and George M. Church. 2008. From genetic privacy to open consent. *Nature Reviews Genetics* 9(5): 406–411.

Monroe, Scott M., F. Samantha, and S.F. Anderson. 2015. Depression: The shroud of heterogeneity. *Current Directions in Psychological Science* 24(3): 227–231.

Morris, Sarah E., and Bruce N. Cuthbert. 2012. Research domain criteria: Cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience* 14: 29–37.

Nissenbaum, Helen. 2004. Privacy as contextual integrity. *Washington Law Review* 79: 119–157.

Owen, Michael J. 2014. New approaches to psychiatric diagnostic classification. *Neuron* 84(3): 564–571.

Poldrack, Russell A., and Krzysztof J. Gorgolewski. 2014. Making big data open: Data sharing in neuroimaging. *Nature Neuroscience* 17(11): 1510–1517.

Serepkaite, Jurate, Zivile Valuckiene, and Eugenijus Gefenas. 2014. 'Mirroring' the ethics of biobanking: What should we learn from the analysis of consent documents. *Science and Engineering Ethics* 20(4): 1079–1093.

Shin, Jaehoon, Guo-li Ming, and Hongjun Song. 2014. Decoding neural transcriptomes and epigenomes via high-throughput sequencing. *Nature Neuroscience* 17(11): 1463–1475.

Song, Chunyao, and Tingjian Ge. 2014. Aroma: A new data protection method with differential privacy and accurate query answering. Proceedings of the 23rd ACM international conference on Conference on Information and Knowledge Management-CIKM '14, 1569–1578. Shanghai: ACM.

Soria-Comas, Jordi, and Josep Domingo-Ferrer. 2015. Co-utile collaborative anonymization of microdata. In *Modeling decisions for artificial intelligence-MDAI 2015*, LNCS 9321. Springer, to appear.

Stonnington, Cynthia M., Tan Geoffrey, Klöppel Stefan, Chu Carlton, Draganski Bogdan, Clifford R. Jack Jr, Chen Kewei, Ashburner John, and Frackowiak Richard SJ. 2008. Interpreting scan data acquired from multiple scanners: A study with Alzheimer's disease. *Neuroimage* 39(3): 1180–1185.

Thagard, Paul. 2008. Mental illness from the perspective of theoretical neuroscience. *Perspectives in Biology and Medicine* 51(3): 335–352.

Thompson, Paul M., et al. 2014. The ENIGMA consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior* 8(2): 153–182.

U.S. Federal Trade Commission. 2014. Data brokers: A call for transparency and accountability. Available at: https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf. Accessed 6 Sept 2015.

U.S. Homeland Security Department. 2009. A roadmap for cybersecurity research. Available at: http://www.dhs.gov/sites/default/files/publications/CSD-DHS-Cybersecurity-Roadmap.pdf. Accessed 6 Sept 2015.

Van den Hoven, Jeroen. 2008. Information technology, privacy, and the protection of personal data. In *Information technology and moral philosophy*, ed. Jeroen Van den Hoven and John Weckert, 301–321. Cambridge: Cambridge University Press.

Walter, Henrik. 2013. The third wave of biological psychiatry. *Frontiers in Psychology* 4: 582.

Walzer, Michael. 1983. *Spheres of justice: A defense of pluralism and equality*. New York: Basic Books.

Wang, Xiao-Jing, and John H. Krystal. 2008. Computational psychiatry. *Neuron* 84(3): 638–654.

Part IV
Ethical Governance

Big Data Governance: Solidarity and the Patient Voice

Simon Woods

Abstract Rare diseases are individually rare but collectively form a population of 30 million people within Europe alone. Most rare diseases are genetic in origin and recent research initiatives are bringing the latest genetic technologies, including whole genome sequencing, together with medical records and natural history data. The rareness of these conditions means that strategies for data sharing are a necessity to ensure that patients are able to obtain a diagnosis and the potential for treatment. Rare disease research is therefore a preeminent example of biomedical "Big Data". This chapter explores the social and ethical challenges of biomedical "Big Data" with a focus on two case studies of contemporary rare disease research and through the framework of "solidarity" as developed by Prainsack and Buyx (2011, 2013). The analysis presented in this chapter is sympathetic to the concept of solidarity as the basis for a governance model for biomedical "Big Data" research. However there are some limitations to the solidarity model and it is argued here that a presumption of solidarity may presume *too* much. The principle of solidarity is very evident within the history of rare disease patient activism but this has evolved alongside other practices, characterised here as "the patient voice" which demands a more collaborative approach to the governance of research. The collaborative approach is one which allows the patient voice to be heard and respected thereby giving research participants an opportunity to be able to negotiate the conditions of participation in research. The chapter concludes with some reflections upon the future challenges for biomedical "Big Data" governance.

1 Introduction

There are up to 30 million people living with a diagnosis of "rare disease" (RD) across Europe. Individually rare but collectively numerous the challenge of rare disease is recognised as a global concern to health providers and policy

S. Woods (✉)
Policy Ethics and Life Sciences Research Centre Claremont Bridge (4th Floor), Newcastle University, Newcastle upon Tyne, NE1 7RU UK
e-mail: simon.woods@newcastle.ac.uk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_10

makers. The low prevalence in individual countries means that international research collaboration is necessary and this entails data sharing and inter-operable research structures. Up to 80 % of rare diseases are genetic diseases and strategies that seek to combine "omics" data with whole genome sequencing data, data from medical records, natural history data and data on family members of the proband (the affected individual) are now regarded as essential research tools. This combination of data sources opens a potential for the exploitable re-purposing of research data and presents the research participant with the challenge of consenting to a complex context of biomedical "Big Data".

There are many rare diseases that are undiagnosed and patients are recruited into research to, amongst other things, achieve a potential diagnosis by taking advantage of the technologies only available through a research project. Seeking a genetic diagnosis is sometimes a motivating element for participation in research because it is a threshold factor for entry into other research, which may lead to potential therapy, and so there is a strong incentive for patients to participate in research in which data sharing is a requirement. Data sharing initiatives of this kind, either implicitly or explicitly, call upon the vested interests of potential, and actual, participants and use these "interests" to provide a particular justificatory framework for governance. These interests usually include two important types, personal or *individual* interests and *common* or *solidarity* interests in a way which could be seen as corresponding to what Prainsack and Buyx (2013) describe as "tier 1 solidarity" (2013, p.75).

Research within RD genomics is a pertinent example of biomedical "Big Data" research that presents a complex case study of the ethical and governance issues of participation. This chapter explores those issues by first setting out the recent scholarship on solidarity, mainly drawing upon the work of Prainsack and Buyx (2011, 2013) before going on to use two examples of ongoing RD genomics research to discuss governance practices through the solidarity framework. By introducing the concept of the "patient voice", a concept originating within the patient activism movement, the chapter goes on to draw out some of the limitations of the solidarity model as well as the potential benefits of combining solidarity practices with principles drawn from patient activism. The chapter concludes with some reflections upon the future challenges for biomedical "Big Data" governance.

2 Background

There are many international research consortia that have taken rare disease (RD) as a focus. The International Rare Diseases Research Consortium (IRDiRC) links researchers and organizations involved in RD research and through these collaborations it seeks to achieve two main objectives by the year 2020: to deliver 200 new therapies for rare diseases, and the means to diagnose most rare diseases. These are significant ambitions and offer the potential of direct benefits to the

many families who are living with undiagnosed and sometimes diagnosed but untreatable RD. To make these goals realistic the consortium is working towards several targets; establishing and providing access to harmonized data and samples; performing the molecular and clinical characterization of rare diseases; boosting translational, preclinical and clinical research streamlining ethical and regulatory procedures (Knoppers 2014; Mascalzoni et al. 2015).

One of the consortium members, and one of the case studies for this chapter, is the European Union funded project RD: Connect. This project is perhaps an archetype of a biomedical "Big Data" project as it seeks to create an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. The platform will enable researchers to access complete clinical profiles combined with -omics data and bio-samples from RD research tissue banks. The processed data from collaborating research projects is held in a database, where it is combined with other omics data types plus phenotypic and biomaterial information. A data access committee must approve researchers wishing to access data via which will be pseudonymized using a unique global identifier (Hansson et al. 2016).

The second example is the UK's "100, 000 Genomes Project" (100 K GP) which has RD as one of its main research interests. The aim of the 100 K GP is to routinize the use of whole genome sequencing and related technologies for both international research and for clinical purposes within the National Health Service (NHS). 50,000 genomes will be obtained in relation to RD but not all the individuals involved will be affected. In order for enough useful information to be obtained the project will adopt the "trios" approach, ideally analysing three genomes per patient (the affected person plus their parents or two blood relatives). Both projects face similar challenges; of making meaningful sense of the masses of data, for example, by distinguishing between harmless natural variations between individuals and variations that are clinically significant. Both RD: Connect and 100 K GP face the challenges of managing the security of the data and of translating the data into direct patient, and wider societal, benefit.

Both projects employ technologies that will generate unequalled amounts of bioinformatics data and will be capable of identifying genetic variants unrelated to the focus of the research or diagnostic intention in addition to variants of unknown significance. The range of research results to emerge from such projects will include pertinent individual findings as well as incidental findings (Wolf et al. 2012). Both will create ethical challengers for the researchers, including how to create pathways to manage such findings and whether to offer them to research participants or their clinicians. There has been considerable ethical debate engendered by the range of approaches adopted regarding the management of findings and though there is no international consensus on how findings ought to be managed there are discernible commonalities of approach that generally require participants to accept restricted options for return of findings. Both projects also face the uncertainty of how the consequences of generating such large volumes of data alongside a commitment to wide data sharing will be managed ethically in the future.

2.1 The Rare Disease Context

Rare disease (RD) patients and their families have particular characteristics that distinguish them from both the general population and those with common diseases. 80 % of rare diseases are genetic and are frequently life-limiting and disabling (Schieppati et al. 2008). Half of the people with RD are diagnosed at birth or in childhood; adding further complexity to the participation in data sharing research. Participation of affected children requires parental consent, parents who are frequently also facing emotional and practical issues (EURORDIS 2015). There are few effective treatments for RD so there is a strong interest in research within the RD community especially where this has the potential to lead to novel treatments; and most research encompasses this possibility within its vision no matter how remote that possibility really is.

The RD community is represented by patient organisations, often formed by parents as a way of providing mutual support and often also to advance research and treatment of RD (Andersen 2012; Woods and McCormack 2013). These interests are manifested by activities such as: funding research directly and indirectly; lobbying for the allocation of institutional and government funding; seeking to influence legislative and structural mechanisms so they are enacted positively to enable research; advocating for research to pursue particular directions; and being directly involved in the governance of research projects (Rabeharisoa 2003, 2006). This culture of patient activism is a significant social factor in the context of the governance of RD research and what it brings to understanding the concept of solidarity in RD research will be explored further below.

On a global level, the advances in genetic technologies have already 'changed the landscape of rare disease research' (Boycott 2013:681), with more than 130 new gene identifications in 2012, greater than a threefold increase on the previous 3 years combined. The founding of the International Rare Diseases Research Consortium (IRDiRC) in 2011 and its ambitious targets for RD diagnostics and research for all rare diseases and 200 therapies by 2020 indicates that RD is to the forefront of developments in the field (IRDiRC 2015). The use of advanced technologies such as whole genome sequencing and technologies enabling data sharing are regarded as absolutely necessary RD and places this area of research to the forefront of biomedical "Big Data" consumers.

3 Solidarity and RD Big Data

Prainsack and Buyx's (2011, 2013) analysis of the concept of solidarity and application to the governance of research biobanks is a valuable framework for the analysis of the two case studies presented in this chapter. In summary the solidarity approach takes as its starting premise the idea that people do, and are willing to, accept costs to themselves in order to assist others for more general

benefits (common goods). They distinguish this principle from that of self-sacrifice or altruism because the intention to help others may also run in parallel with the hope of a personal benefit of some kind. Put into the context of research biobank participants may see the potential benefit to others with particular health problems but also see themselves within the mass of others who make up that population. The model may similarly be extended to the RD context where, despite the more pressing desire to see direct personal/family benefit, there is also a strong desire to see benefit to others at some personal cost to the self (Woods and McCormack 2013). In addition RD research participants bring another strand to the solidarity discourse, namely their desire to negotiate the terms of governance and to set boundaries to the kinds of costs such complex research projects can reasonably impose. This perspective will be referred to as *the patient voice.*

3.1 Solidarity: A Moral Concept?

Prainsack and Buyx do not see solidarity as a political or moral concept and nor do they attempt to offer a normative justification for solidarity. They are careful to emphasise that their account of solidarity is of a set of practices, which may be discerned within three levels of institutionalisation. At tier 1, or the interpersonal level, informal practices of solidarity emerge out of a sense of similarity or shared causes. At tier 2 practices of solidarity, though still informal, become routinized within a given group or community. Tier 3 solidarity is where regulation or law mandates these practices. The direction of evolution of solidarity practices is not of necessity "bottom up" but may in principle be "top down" (2013:76).

Before moving on to consider the case studies in the light of the solidarity framework I will first comment on the claimed moral neutrality of solidarity. At a descriptive level it is clear that that solidarity practices are not *necessarily* moral practices. Though it is important to note that the logical form of solidarity as a *practice* implies that solidarity is something that requires a community of practitioners, who endorse and promulgate those practices. As a set of shared and endorsed practices solidarity is therefore quite clearly axiological in nature, the kind of practice to be promoted as a form of good, or at least as a means to achieving a form of good to which there is, at a minimal level, a consensus that the practices are worthwhile and ought to be adopted. In describing a solidarity-based approach to biobank governance Prainsack and Buyx have focused upon a form of human activity, at least in the context of public biobanks, that is frequently described as being a particular form of benevolent human endeavour and not infrequently described as an endeavour of moral importance (Solbakk et al. 2009; Kaye et al. 2012a, b). In their own words willingness "to contribute to research that assists and benefits others" (2013, p.76) is surely more than a description of mere co-operation towards some goal? Their meaning, if not their language, is that of beneficence and beneficence *is* a moral practice.

They go on to a say that "solidarity should underpin the relationship between participant and biobank, alongside autonomy" (2013, p.77) presumably because they believe that solidarity has moral work to do as leverage against the ubiquitous moral principle of respect for autonomy. One of the reasons they place solidarity in contrast to autonomy is presumably because they wish to create a space for a range of governance practices that are not merely nailed down to practices of informed consent; the "quasi-synonym for autonomy itself" (2013, p.78). Thus the implication of the solidarity-based approach is that it will bring a revision that amounts to a *better* form of governance for biobanks.

The solidarity premise, the willingness to accept a personal cost for the benefit of others, is emphasised by Prainsack and Buyx as a potential starting point for governance structures. They contrast the solidarity premise to the risk-based approaches to governance, which they argue have dominated recent research governance. They also distinguish solidarity-based governance in terms of the relative weight it ascribes to notions of individual autonomy. In their account, autonomy is a concept associated with highly bureaucratised approaches to informed consent where consent is made to do the bulk of the ethical work within research, mitigating risk and offsetting potential litigation against the researchers. However there is a risk that accepting the legitimacy of governance practices on the basis of solidarity may be too permissive; especially when the governance structure is "top down". For example the tendency for biobank, and to an extent, data research to streamline consent along the lines of broad and permissive approaches has become highly routinized (Steinsbekk et al. 2013). This routinization can be evaluated from two perspectives, one is that it is a "good" example of solidarity in practice, bothering less with the detailed bureaucracy of informed consent practices and allowing researchers to get on with the task in hand (Cassell and Young 2002). The other perspective is that it is convenient short-cut for researchers to pursue their own interests under the wide mantle of presumed solidarity, a seemingly cynical comment but not without corroboration. What the patient voice perspective brings to these debates is twofold: first, the linear structure of tiered solidarity is disrupted, tier 1 "bottom up" solidarity can be curved around, so to speak, to meet tier 3 "top down solidarity" through the active involvement of the patient voice. Second the detail of governance practices can be negotiated in such a way that claims to solidarity as an underpinning value are negotiated and clarified in ways that do not conflict with other important interests.

4 Solidarity and Biobanks

Prainsack and Buyx's example of solidarity practices are focussed upon the context of public research biobanks. They refer to the example of the UK Biobank as a public health biobank that aims to create a rich database from which to research serious conditions of middle and old age, including cancer, diabetes, heart disease and dementia (Wellcome Trust 2015). Although the UK Biobank does not explicitly

utilise the concept of solidarity it is one that, according to Prainsack and Buyx "resonates" well with the solidarity approach (2013: p: 85).

What does such an approach entail? Certainly the 'promissory goods' of such a biobanking enterprise have broad appeal to publics who value health and the health benefits which may stem from research biobanks. In solidarity terms the potential benefits to be achieved draw upon the universal appeal of addressing the commonest causes of mortality and morbidity; cancer; heart disease and dementia. Though participants are never promised direct benefits the framing of the enterprise means that it is not difficult for participants to foresee the potential for direct benefits to others and to self. This discourse of beneficence also draws upon a cultural context, which acts as a bulwark to the purpose of research biobanks. For example there is also a quite reasonable expectation that, if benefits should arise, they will be equitably distributed because the UK Biobank and its associated research programmes are located within, or closely allied to, a culture of social health care. In the UK the National Health Service (NHS) in particular has enduring public standing. By adopting this kind of strategy, mass co-operation involving some personal cost, but towards achievable common goods, the UK Biobank has endeavoured to make the case that those who participate in the research will, even under conditions of uncertainty, be making a decision of "well placed trust" (O'Neill 2002).

Using the solidarity model described above the UK Biobank might plausibly be categorised as a tier 3 solidarity institution, with governance procedures that are mainly directed from the top down. This is because under the umbrella of a broad statement of values it exploits the maximum legal powers available to researchers to utilise data and tissues. Participation in the UK Biobank is of course based on informed consent but the consent is broad, allowing for the possibility of an indeterminate research agenda. Participation also comes with non-negotiable conditions; for example with regard to individual feedback where, like many other similar projects, individual feedback is strictly limited. Participants do have some power to negotiate the terms of their involvement by withdrawing their co-operation; but even these are subject to conditions that limit the possibility of total withdrawal of data and tissues.

As a research organisation the UK Biobank follows a long line of similar medical research institutions, which, from a governance perspective, might as easily be described as governance by 'elites'. It is elites who set both the research agenda and determine the conditions under which the co-operation of participants will be canvassed and managed. The 'governance by elites' model may still employ the rhetoric of solidarity, health related research often does this, but as the tissue retention scandals of Bristol and Alder Hey (Redfern Report 2001) revealed, the liberties presumed by the elites, once exposed to wider public scrutiny, can be found severely wanting. However, this is not a charge that can be made against the UK Biobank; through its own resolve to foster public debate and demonstrate transparent governance practices, it has placed itself as far removed from the discredited research regimes that caused the health research scandals. Nevertheless these scandals have had a formative influence on biobank research governance. The

UK Biobank, for example, operates under a reformed legal structure, which includes the Human Tissue Act (2004) and the oversight of the Human Tissue Authority which has the power to remove an institution's licence and can criminally prosecute violations of the Act. In addition, the UK Biobanks' approach to governance exudes awareness of the need for accountability, trust and transparency that have, to a considerable extent, been driven by public concern, legal challenge, as well as critical academic scrutiny of the evolution of biotechnology in the post-genomic era (Dresser 2001; Nightingale and Martin 2004; Tutton and Corrigan 2004).

The UK Biobank continues to foster public dialogue and canvasses ethics and governance guidance from an especially convened Ethics and Governance Council who act as a 'critical friend'. As a research organisation it has been under considerable critical academic scrutiny (Levitt and Weldon 2005; Kaye et al. 2012b) and through its willingness to support such scrutiny and evaluation it has become something of a role model for similar research biobank projects. The range of concerns, which shaped the UK Biobank model can also be seen as influencing the way that Genomics England (GE) has established its governance procedures for the 100 K GP. Similarities include trading upon the willingness of individuals to accept some personal cost for the benefit of others, the potential for wider social goods including national economic goods all couched in the language of reassurance that the trust of participants will be well placed. A quite noticeable strategy employed by GE, evident within its web-pages, is the placing of the 100 K GP within an historical context of successful and pioneering medical science. The "Time-line" section places 100 K GP as downstream from the discovery of DNA and the sequencing of the first human genome. Other contributors quoted on its web-pages draw upon a much wider history; Simon Stevens, NHS England's Chief Executive is quoted as saying:

> The NHS is now set to become one of the world's 'go-to' health services for the development of innovative genomic tests and patient treatments, building on our long track record as the nation that brought humanity antibiotics, vaccines, modern nursing, hip replacements, IVF, CT scanners, and breakthrough discoveries from the circulation of blood to the existence of DNA . . . the NHS' comparative advantage in unlocking patient benefits from the new genomic revolution stems from our unique combination of a large and diverse population, with universal access to care, multi-year data that spans care settings, world-class medicine and science, and an NHS funding system that enables upstream investment in prevention and new ways of working, as demonstrated by this ground-breaking 100,000 Genomes Project. (Genomics England 2014 "News" http://www.genomicsengland.co.uk/uk-to-become-world-number-one-in-dna-testing-with-plan-to-revolutionise-fight-against-cancer-and-rare-diseases/)

Another similarity between the UK Biobank and Genomics England is that both were established in the wake of major embarrassment for the Department of Health that had dented public trust in the NHS. The UK Biobank was launched a few years following the publication of the Redfern Report (2001) into the Alder Hey tissue retention debacle. GE's up-beat declarations of the strong heritage and aspirational future of health care in the UK came only months after the Government's publication of "Hard Truths: The Journey to Putting Patients First"

(Department of Health (DH) 2014) in which they responded to the public enquiry into the serious failings in care at Mid-Staffordshire NHS Foundation Trust before 2009. The response includes a re-affirmation of key values, for example to "put patients first" and a pledge to listen to the vulnerable (DH 2014, p.6).

It will be an interesting task to observe how the 100 K GP, a major and intrusive research project, which unlike UK Biobank, seeks to recruit patients with particular clinical conditions including cancer and rare diseases, goes about practicing its own style of governance. Like the UK Biobank, GE also employs a top down governance structure, as they state: "Genomics England is following established legal and regulatory standards to seek consent of potential participants" (Genomics England FAQ). Genomics England, though nominally a private company, has direct UK Government endorsement. Jeremy Hunt, the Secretary of State for Health, is quoted on the GE website: "The NHS has a long track record as a leader in medical science advances and it must continue to push the boundaries by unlocking the power of DNA data."

With the Secretary of state for Health being the sole share-holder, GE is a flagship Government project and became operational immediately prior to the General election of 2015. As they state on the web pages: "Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy, through the sequencing of 100,000 genomes." This kind of rhetoric, rallying the public to a common cause, is redolent of solidarity and implicitly calls upon the duty of citizens to endorse and potentially participate in such enterprise. The same rhetoric has been employed in relation to genetics (Department of Health 2003) and other biotechnological initiatives that have high-level government support including stem-cell science (Department of Health 2011) and more recently regenerative medicine:

> Regenerative medicine has the potential to play an increasingly vital role in delivering the next generation of healthcare, offering treatments or possible cures for areas of unmet medical need such as Parkinson's, diabetes, stroke and heart disease. There is also a real possibility for us here in the UK to gain economic benefit from commercially exploiting this exciting area of technology. (Department of Health: Ministerial Foreword 2011, p.3)

Plows places a slightly more cynical gloss on similar such claims: "The promissory discourses of the bio-economy are a form of economic 'futures trading'; invest in this and there will be economic and health rewards" (2010: p.168). It is interesting to observe that the language employed by the promoters of these national research enterprises does fluctuate between economic and moral investment and both will offer good returns. As Nightingale and Martin observed over a decade ago: "many expectations are wildly optimistic and over-estimate the speed and extent of the impact of biotechnology" (2004, p.564). When governments advance research agenda that are so 'obviously' in a nation's common interest then the opportunity for slippage to occur within a solidarity discourse is real. The recognition that people are in principle willing to pay a cost for the benefit of others can be seen by Governments as grounds to move from a *reasonable expectation* to a justified lack of consultation; as in the case of Iceland's genome project discussed below.

The UK Government's stance in relation to the 100 K GP is perhaps more fairly characterised as located at the strong end of "reasonable expectation" perhaps because of the salutary lessons learned from less cautious projects. However the combining of different kinds of "common good" has the potential to devalue the virtue of solidarity. Placing specific health benefits anticipated from the project alongside more generic, possible economic, benefits blurs the status of the participant. The potential here is to shift the regard for the citizen as a volunteer solidarist, taking on certain costs for related benefits to self and others, to seeing the citizen as an economic commodity.

One of the more notorious examples of top down governance in which something like a presumption of a duty to participate occurred was the Icelandic Government's collaboration with deCode Genetics. The decision to allow a private biotech company exclusive access to the Icelandic population's health data was sweetened with the promise of specific health and economic benefits. The subsequent "rebellion" in terms of critical international debate, public protest, and professional non-compliance brought the project to a premature end (Árnason and Andersen 2013). Amongst the concerns raised in these debates regarded the potential for a private company to profit from what was being presented as a project of solidarity for the common good.

A similar response, though at a more modest scale, occurred within the UK over the attempt by the Government to implement care.data, a scheme to enable wide health related data sharing, with the promise of a more efficient use of data for clinical care and beneficial research. The implementation of care.data was mishandled by the Department of Health, with inadequate public information, and an aggressively critical media campaign, which exposed practices such as data-sharing with the insurance industry, widely perceived as a violation of trust. One of the concerns expressed was the sharing of data with private companies who may be able to exploit the data for profit and without a reasonable benefit to the NHS. There is still an intention to roll out the care.data programme but it is now being conducted in a much more precautionary way, piloting data sharing at a number of sites and canvassing public opinion as the programme develops; though there is widespread concern that the initiative will never recover from the initial loss of trust (Little 2015).

DeCode Genetics and care.data, in different ways, illustrate one of the challenges of adopting a tier 3 solidarity approach with its attendant "top down" processes of implementation. Even for projects that are likely to deliver the general good they claim, there is still the problem of what political philosophers call "endorsement constraint" (Kymlicka 1990). Endorsement constraint means that an assertion that policy X is good is not translated into a belief (an endorsement) that policy X is good. Adopting a strategy of imposing policy X may become high risk for a Government, may not deliver the improvements promised and, as in the case of Iceland with deCode Genetics, and to a degree the UK with care.data, may do actual harm especially in relation to public trust.

How does a Government go forward in implementing a strategy that calls upon solidarity, requires the co-operation of its citizens more broadly, or requires the particular co-operation of certain sectors of society? The 100 K GP is an interesting

contemporary example of such an approach. Genomics England (GE) have gone about actively identifying the nature of the problem to be faced, have set out a strategy for meeting the challenges, identified the benefits that will follow and reassured the public of good governance and their best endeavours to adhere to the highest ethical standards. In addition, as with UK Biobank, it is a project that draws upon the regard held by most UK citizens for the work of the NHS.

GE, in fairness, have gone much further than merely asserting their intentions, they have commissioned qualitative research to inform the development of the information and consent procedures, they have actively canvassed critical discussion of their programme and appointed a professional science media presenter to head the engagement and involvement strategy. Like the UK Biobank they have established an independent Ethics Advisory Committee and established an active website which provides public access to news, reports and governance documents. Through these and other techniques GE has tried to establish a conversation with the general public in which at least a seemingly open exchange of ideas can occur. However what GE asks of participants is committing and open ended; as can be seen from this extract from the RD participant information document:

- You are agreeing that past medical records (from birth), as well as current and future information about your condition (health data) can be collected by the 100,000 Genomes Project.
- You are agreeing that this information can be studied now and after your death.

The 100 K GP has an ambiguous position as a research biobank, though it is a private company it is Government controlled and through that status seeks to offer reassurance about the standard of governance to be expected, implicitly, as accountable as any Government department. Though it is a research project it is not only a biobank or database but also part of a strategy of clinical implementation of genomics across the NHS. Though the benefits are potentially there for all users of (initially) NHS England the immediate costs of participation are born by particular groups within the population, people with or related to people with RD and cancer. As the GE website puts it: "The UK will become the first ever country to introduce this technology in its mainstream health system – leading the global race for better tests, better drugs and above all better, more personalised care to save lives."

An example of how GE have encouraged partnerships and established mechanisms for informing their governance approaches is the commissioning of Genetic Alliance UK to canvass opinion on genomic technologies. Genetic Alliance UK is a charitable organisation which seeks to support people living with genetic conditions as well as to increase public and government awareness of the impact of living with such conditions; it is very much within the tradition of organisations who foster the "patient voice". Genetic Alliance UK conducted an online survey inviting patients and family members affected by suspected or confirmed rare genetic conditions to take part. The survey received 231 responses and a small number of qualitative interviews were conducted by telephone. The survey and interviews explored patients' and families' views around several key areas including consent and the management of findings. Genetic Alliance UK published their results in the form of a patient's charter (Genetic Alliance UK 2015).

Regarding the management of findings the charter states: "The majority of patients want to know anything that a geneticist might accidentally discover during the analysis of their genome, regardless of the seriousness of the condition the finding relates to, or whether it is clinically actionable" (2015, p.11).

In terms of what should be looked for: "Patients told us that they would want the person analysing their genome to purposefully look for genetic alterations that have been linked to conditions that are unrelated to the original diagnostic aim. The seriousness of the condition being searched for as well as clinical outcome, did not greatly affect patients' views on this issue." (2015, p.14)

However the GE policy takes a much more controlled and precautionary approach with only an option to receive a very limited list of additional or secondary findings that are both known to cause serious conditions and are actionable (e.g. certain types of familial bowel cancer and the breast cancer genes; BRCA1/2 genes). Incidental findings will not be looked for or reported on in an approach that echoes other recommended strategies, though is still more restrictive (ACMG 2013, 2015). Genetic Alliance UK do note in their charter that, while supporting maximal choice for participants, they also acknowledge that there should be no "one size fits all" because RD families are heterogeneous when it comes to decisions about genetic information. This is presumably a factor which has influenced their position on consent which states that: "patients indicated that dynamic consent would be the most appropriate model for consent in the context of genome sequencing in the NHS for two reasons. Firstly, the majority of respondents thought it was important for patients to be able to change their mind about what information they want to receive from genome sequencing. Secondly, a significant proportion of respondents felt that the individual should be able to determine what results they are given following the analysis of their genome" (2015, p.17). The only dynamic aspect to consent within GE policy however is with regard to optional additional findings.

It is very clear that GE place a lot of emphasis on the consensual nature of participation within the 100 K GP as part of their robustly ethical approach. The adoption of a static broad consent is in keeping with similar projects in which there are good grounds to accept that broad consent is an ethical form of informed consent (Sheehan 2011). However it is significant to note that a project that encourages participation and engagement does not seem to be open to negotiation where the views of potential participants differ significantly from that standard model (Hoeyer 2010). These comments are made in the early days and how the project evolves will no doubt be keenly observed.

5 Solidarity and Patient Activism

A different process in the evolution of governance strategies, one that sits closer to tier 1 solidarity, 'bottom up' practices can be discerned within the contemporary history of citizen activism. There are several strands to these solidarity practices;

some related to political consumerism and some to civil rights. Contemporary activism is often a response to causes related to the provision and governance of social goods, equality, health and education. Women's and disability rights groups in particular contested the quality of, and access to, health care, challenged medical paternalism and insisted upon the right to participate in healthcare decisions made about them (Lupton et al. 1991; Rodwin 1994; Ruzek 2007). Activism that focussed substantively on matters of health and health care provision were early indicators of an emerging phenomenon characterised by later commentators as "biological citizenship" (Rose and Novas 2005), except that the obligations of the citizen are reconfigured as consumer or citizens' rights. The consumer analogy is poignant to the debates elaborated in this chapter, as has been illustrated with the examples above, the discourses around calls for citizen support of burgeoning biotechnology research draws upon the language of both citizenship and economics. As Waldby (2011) has argued citizens are not just citizen contributors to socially good causes but can also be seen as economic actors and sometimes as commodities within the complex bioeconomy of medical research. Human tissue, genetic information, and access to cohorts of research participants have become the basis for a kind of commerce and at times the material of negotiation, between patients and professionals (Woods and McCormack 2013).

An early example of bio-activism was in the context of HIV/AIDS research. Collective action influenced clinical research by challenging the restrictions on access to potential therapies, questioning the validity of standard research methods such as the randomized controlled trial and challenged pharmaceutical regulation (Epstein 1995). HIV/AIDS activism created a role model including for RD groups such as the French Muscular Dystrophy Association AFM –Telethon founded in 1958 (Houyez 2004). Strategies adopted by patient organizations include learning the language of biomedicine in order to be treated on more equal terms by researchers. This was one of the more literal origins of the concept of the *patient voice.* Activism also involved developing a political voice and involved political lobbying, challenging the regulation of research and fund-raising to establish patient organisations as research paymasters (as happened with AFM) (Rabeharisoa 2003, 2006). The activism approach to research governance is interesting from the perspective of solidarity because there are at least three distinct features. One is that activists do not regard themselves as vulnerable in the conventional sense that the Declaration of Helsinki positions the research participant (WMA 2013). The second is that they construe research participation as a right. The implications of this kind of activism cut across the three tiers of solidarity outlined by Prainsack and Buyx, RD patients are not mere tier 1 solidarists. The third feature is that for RD patients, and family members, their stake in the enterprise of research is a much stronger one than that of healthy volunteers in public epidemiological biobank research; their need is real and present and their aspirations for research to produce benefits is more immediate. This sense of imperative brings a pressure from "below" to make things happen and to move things forward (Aymé et al. 2008). The result of this activism has been the evolution of a collaborative model of managing

research, the "partnership model" that allows the patient voice to be heard (Callon and Rabeharisoa 2003; Rabeharisoa 2003) and which can be seen to be operating within several international research projects including the second case study in this chapter: RD: Connect.

6 RD: Connect: Solidarity and the Patient Voice

RD: Connect, within the IRDiRC consortium has drawn upon a very long heritage of international collaboration in RD research. One of the key partners of RD: Connect is EURORDIS – European Organisation for Rare Diseases European Rare Diseases. EURORDIS is a non-governmental organisation that has formed an alliance of RD patient organisations. EURORDIS was founded in 1997 with support from AFM and now see themselves as one of the leading organisations for the patient voice within Europe (Mavris and Le Cam 2012). Their vision for the patient is of an empowered actor and many of their activities (running summer schools for patients on research methods, providing distance learning, training patient representatives in committee membership) are designed to achieve this. EURORDIS as a partner within the RD: Connect project are responsible for ensuring that patient representatives are invited to meetings and consultations, supporting the learning needs of individuals going forward as patient representatives. In addition they are responsible for canvassing the views of the patient community on key issues arising within the project. The Rare Disease Patient and Ethics Council (RD-PEC) respond to ethical, social and participatory arising from the research taking place in the context of RD-Connect and its allied research projects. The PEC provides an open forum for any interested individual or organisation to ask a question with an ethical focus. In addition there is a Patient Advisory Council (PAC), which acts as a conduit for patient/participant opinion on wider issues to reach the highest level of management within the project. As RD: Connect continues towards the conclusion of its funding in 2018 there will be on-going empirical work within the project; canvassing patient opinion on key issues of concern. This collaborative approach has been utilised in developing important policies for governance within the project including policies on data sharing and consent (Mascalzoni 2014). Patient representatives are also present within the executive groups of RD: Connect providing an opportunity for the patient voice to be heard throughout the project.

There is a strong belief held by RD patient representatives that it is the participants who should be more actively determining the conditions of participation in research. Early RD collaborations have shown that this can be practically possible. An example of this is the TREAT-NMD Global Database Oversight Committee (TGDOC). The Global Registry is a database for neuromuscular diseases and drawing data from national disease registries. The Global Registry ensures that patients who register in their national registry can be contacted if their profile fits the criteria for a clinical trial and is thus regarded as a major resource for facilitating further RD research. The oversight committee is composed of

representatives drawn from across the TREAT-NMD Alliance (www.treat-nmd.eu) including international patient organisations. The oversight committee are invited to vote on the appropriateness of any enquiry made by researchers to the Global Registry. The collaborative practices, consent and data sharing policies, charters and statements of values, evident within projects like TREAT:NMD and RD: Connect are important exemplars of where solidarity meets the patient voice. Work within the RD: Connect group (McCormack et al. 2016) reveals that patient representatives are very cautious about trying to represent the range of concerns and approaches held by patients. Further, they are very willing to act on the principle of solidarity but also value their autonomy and expect governance practices to reflect these concerns by utilising techniques such as dynamic consent, and providing flexibility regarding rights to information and access to feedback.

However, it must be acknowledged that governance practices within projects like RD: Connect and 100 K GP are not, and in some instances cannot, be the product of negotiations within the project between the stakeholders. National and international legislation mandates particular approaches to consent, data access and confidentiality for obtaining, re-using and sharing tissue samples required for the generation of omics data (Parker et al. 2004; Kaye et al. 2012a, b). Patient representatives are not always content with the degree of permissiveness given to researchers to access, use and share tissues and data without consulting or re-consenting patients. Equally they are sometimes discontent with the restrictions law and regulation place upon access to e.g. archive samples of RD tissues, regarding these as potentially wasting a precious resource (Bathe and McGuire 2009).

7 Concluding Remarks

This chapter has explored some of the governance challenges in biomedical "Big Data" from the particular perspective of RD research. RD are by definition rare but collectively they are numerous as is borne out by the estimated 30 plus million people living with rare disease in Europe. Advocates for RD, such as the French Association AFM and EURORDIS, point out that RD is not the problem of a few individuals and their families but is of major social concern like many other threats to health and welfare the world faces. Early campaigning by RD activists reveals that in addition to a call for solidarity amongst RD patients and wider society; they also claim a right for the patient voice to be heard. Drawing upon the work of Prainsack and Buyx (2011, 2013) the concept of solidarity has been explored in the context of two RD research case studies.

The case studies described in this chapter are both at the forefront of biomedical 'Big Data' and both have developed governance strategies to secure, safeguard and protect participants and their data. These governance strategies have been developed cognisant of the need to balance protection and safety of participants against the imperative to ensure maximum beneficial use of the data generated. It is evident that these case studies, like many others in the contemporary context

of biomedical "Big Data" have utilised a combination of governance strategies combining elements of traditional governance and collaborative approaches to developing governance strategies within legal frameworks. It has been argued that the concept of solidarity combined with collaborative approaches, characterised here as the *patient* voice, offers a robust and ethical way forward for international research collaborations on the scale the biomedical 'Big Data' requires. Such projects face an open and uncertain future considering the volume of data likely to be generated, the imperative to facilitate wide data sharing, and the challenges of managing and utilising such volumes of data (Mittelstadt and Floridi 2016). It is likely that these ambitions may only be realised by drawing increasingly upon the resources and experiences of private sector organisations like Google and Apple creating a new context in which the cultures of medical ethics, patient activism meet consumer culture.

Acknowledgment Simon Woods has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 305444 (RD-Connect).

References

American College of Medical Genetics. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine* 15(7): 565–578.

American College of Medical Genetics. 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.

Andersen, T. 2012. The political empowerment of rare disease patient advocates both at EU and national level. *Orphanet Journal of Rare Diseases* 7(2): 1–3.

Árnason, E. and B. Andersen. 2013. deCODE and Iceland: A critique. eLS. doi: 10.1002/9780470015902.a0005180.pub20. Published online: 15 FEB 2013.

Aymé, S., A. Kole, and S. Groft. 2008. Empowerment of patients: Lessons from the rare diseases community. *Lancet* 371(9629): 2048–2051.

Bathe, O.F., and A.L. McGuire. 2009. The ethical use of existing samples for genome research. *Genetics in Medicine* 11: 712–715.

Boycott, K.M., M.R. Vanstone, D.E. Bulman, and A.E. Mackenzie. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 14: 681–691. doi:10.1038/nrg3555.

Callon, M., and V. Rabeharisoa. 2003. Research "in the wild" and the shaping of new social identities. *Technology in Society* 25(2):193–204.

Cassell, J., and A. Young. 2002. Why we should not seek individual informed consent for participation in health services research. *Journal of Medical Ethics* 28(5): 313–317. doi:10.1136/jme.28.5.313.

Department of Health. 2003. *Our inheritance our future*. London: The Stationery Office.

Department of Health. 2011. *Taking stock of regenerative medicine*. London: The Stationery Office.

Department of Health. 2014. *The government response to the Mid Staffordshire NHS foundation trust public inquiry*. London: The Stationery Office.

Dresser, R. 2001. *When science offers salvation. Patient advocacy and ethics*. Oxford: Oxford University Press.

EURORDIS. 2015. http://www.eurordis.org/living-with-a-rare-disease

Epstein, S. 1995. The construction of lay expertise – AIDS activism and the forging of credibility in the reform of clinical trials. *Science, Technology & Human Values* 20: 408–437.
Genetic Alliance UK. 2015. *Genome sequencing: what do patients think? patient charter*. London: Genetic Alliance UK.
Hansson M. G, H. Lochmüller, O. Riess, F. Schaefer, M. Orth, Y. Rubinstein, C. Molster, H. Dawkins, D. Taruscio, M. Posada, S. Woods. 2016. The risk of re-identification versus the need to identify individuals in rare disease research. *European Journal of Human Genetics* 1–6.
Houyez, F. 2004. Active involvement of patients in drug research, evaluation, and commercialization: European perspective. *The Journal of Ambulatory Care Management* 27(2): 139–145.
Hoeyer, K. 2010. Donors perceptions of consent to and feedback from biobank research: Time to acknowledge diversity? *Public Health Genomics* 13: 345–352.
International Rare Disease Research Consortium (IRDiRC). 2015. http://www.irdirc.org/goals/ Accessed 28 Oct 2015.
Kaye, J., L. Curren, N. Anderson, K. Edwards, S.M. Fullerton, et al. 2012a. From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews Genetics* 13: 371–376.
Kaye, J., S.M.C. Gibbons, C. Heeney, M. Parker, and A. Smart. 2012b. *Governing biobanks: Understanding the interplay between law and practice*. London: Bloomsbury.
Knoppers, B.M. 2014. International ethics harmonization and the global alliance for genomics and health. *Genome Medicine* 6(2): 13.
Kymlicka, W. 1990. *Contemporary political philosophy: an introduction*. Oxford: Clarendon Press the University of Michigan.
Levitt, M., and S. Weldon. 2005. A well placed trust? public perceptions of the governance of DNA databases. *Crit Public Health* 15(4): 311–321. doi:10.1080/09581590500523186.
Little L. 2015. Care.data loose ends need tying up now. Opinion. *Health Service Journal*. http://www.hsj.co.uk/comment/caredata-loose-ends-need-tying-up-now/5085349.article Accessed 12 Oct 2015.
Lupton, D., et al. 1991. Caveat emptor or blissful ignorance? Patients and the consumerist ethos. *Social Science and Medicine* 33: 559–568.
Mascalzoni D, E. Dove, Y. Rubinstein, H. Dawkins, A. Kole, P. McCormack, S. Woods, O. Riess, F. Schaefer, H. Lochmüller, B. Knoppers, M. Hansson. 2014. International charter of principles for sharing bio-specimens and data. *European Journal of Human Genetics*. 23(6): 721–728.
Mascalzoni, D., E. Dove, Y. Rubinstein, H. Dawkins, A. Kole, P. McCormack, S. Woods, O. Riess, F. Schaefer, H. Lochmüller, B. Knoppers, and M. Hansson. 2015. International charter of principles for sharing bio-specimens and data. *European Journal of Human Genetics* 23: 721–728. doi:10.1038/ejhg.2014.197.
Mavris, M., and Y. Le Cam. 2012. Involvement of patient organisations in research and development of orphan drugs for rare diseases in Europe. *Molecular Syndromology* 3(5): 237–243. doi:10.1159/000342758.
McCormack, P, A. Kole, S. Gainotti, D. Mascalzoni, C. Molster, H. Lochmüller, S. Woods. 2016. "You should at least ask". The views of rare disease patients and advocates on large scale systems for data and biosample sharing. *European Journal of Human Genetics*. doi:10.1038/ejhg.2016.30.
Mittelstadt, B.D., and L. Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.
Nightingale, P., and P. Martin. 2004. The myth of the biotech revolution. *Trends in Biotechnology* 22(11): 564–569.
O'Neill, O. 2002. *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.
Parker, M., R. Ashcroft, A.O.M. Wilkie, and A. Kent. 2004. Ethical review of research into rare genetic disorders. *BMJ* 329: 288.
Plows, A. 2010. *Debating human genetics: Contemporary issues in public policy and ethics*. New York: Routledge.

Prainsack, B., and A. Buyx. 2011. *Solidarity: reflections on an emerging concept in bioethics*. Swindon: Nuffield Council on Bioethics.

Prainsack, B., and A. Buyx. 2013. A solidarity-based approach to the governance of research biobanks. *Medical Law Review* 21(1): 71–91. doi:10.1093/medlaw/fws040.

Rabeharisoa, V. 2003. The struggle against neuromuscular diseases in France and the emergence of the "partnership model" of patient organisation. *Social Science and Medicine* 57: 2127–2136.

Rabeharisoa, V. 2006. From representation to mediation: The shaping of collective mobilization on muscular dystrophy in France. *Social Science and Medicine* 62: 564–576.

Redfern Report. 2001. The report of The Royal Liverpool Children's Inquiry. London: The Stationery Office. http://www.rlcinquiry.org.uk/

Rodwin, M.A. 1994. Patient accountability and quality of care: lessons from medical consumerism and the patients' rights, women's health and disability rights movements. *Am J Law Med* 20: 147–167.

Rose, N., and C. Novas. 2005. Biological citizenship. In *Global assemblages: Technology, politics and ethics as anthropological problems*, ed. A. Ong and S.J. Collier, 439–463. Oxford: Blackwell.

Ruzek, S. 2007. Transforming doctor-patient relationships. *Journal of Health Services Research & Policy* 12: 181–182.

Schieppati, A., et al. 2008. Why rare diseases are an important medical and social issue. *Lancet* 371(9629): 2039–2041.

Sheehan, M. 2011. Can broad consent be informed consent? *Public Health Ethics* 4(3): 226–235.

Solbakk, J.H., S. Holm, and B. Hofmann. 2009. *The ethics of research biobanking*. Dordrecht: Springer.

Steinsbekk, K.S, B. Kåre, K. Myskja, B. Solberg. 2013. Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? *European Journal of Human Genetics* 21:897–902. doi:10.1038/ejhg.2012.282; published online 9 Jan 2013.

TREAT-NMD Global Database Oversight Committee (TGDOC). http://www.treat-nmd.eu/resources/patient-registries/global-registries/governance

Tutton, R., and O. Corrigan. 2004. *Genetic databases: Socio-ethical issues in the collection and use of DNA*. London: Routledge.

Waldby C (2011) Citizenship, labor and the biopolitics of the bioeconomy: Recruiting female tissue donors for stem-cell research. *Scholar & Feminist Online* Spring 9.1/9.2: special double issue Critical Conceptions: 9 Technology, Justice, and the Global Reproductive Market.

Wellcome Trust: UK Biobank. 2015. Accessed at: http://www.wellcome.ac.uk/Funding/Biomedical-science/Funded-projects/Major-initiatives/UK-Biobank/

Wolf, S.M., B.N. Crock, B. Van Ness, et al. 2012. Managing incidental findings and research results in genomic research involving biobanks & archived datasets. *Genetics in Medicine: Official Journal of the American College of Medical* 14(4): 361–384. doi:10.1038/gim.2012.23.

Woods, S., and P. McCormack. 2013. Disputing the ethics of research: The challenge from bioethics and patient activism to the interpretation of the declaration of Helsinki in clinical trials. *Bioethics* 27(5): 243–250.

World Medical Association Declaration of Helsinki. 2013. *Ethical principles for medical research involving human subjects*. Ferney-Voltaire: World Medical Association.

Premises for Clinical Genetics Data Governance: Grappling with Diverse Value Logics

Polyxeni Vassilakopoulou, Espen Skorve, and Margunn Aanestad

Abstract This chapter discusses emerging tensions related to data ownership and sharing in global genetic data repositories, accessed by researchers and clinicians, from both public and private institutions. We describe the on-going controversies around collecting and sharing genetic mutation data on the *BRCA1* and *BRCA2* genes: the creation of the Breast Information Core (BIC) database in 1995, the decision by Myriad Genetics to stop sharing information in 2004, the subsequent reaction from the community through the "Sharing Clinical Reports Project" and "Free the Data" initiatives and the recent creation of the open ClinVar repository and the public-private BRCA Share resource. We analyse these experiences, elaborate on the unique characteristics of BRCA data and identify different logics shaping the field. Based on this analysis we suggest drawing from the literature on collective action and the governance of commons for addressing the multiplicity of logics identified. We conclude by discussing the usefulness of foregrounding fundamental questions related to equity, efficiency and sustainability for shaping and evaluating governance arrangements in the field.

1 Introduction

The on-going blurring of boundaries between research and clinical care coupled with the possibilities brought-in by advancements in gene sequencing and computational technologies pose challenges to the governance of genetic data repositories. The need to establish data governance arrangements is becoming now pressing because the volume of genetic data is growing exponentially as genetic tests for hereditary conditions are becoming less costly, private laboratories are expanding their menus of diagnostic and prognostic tests, and a growing business market for physician-ordered and direct-to-consumer testing is created. The flux within the field creates ethical tensions that go beyond much discussed privacy and security issues. A recent report from the Nuffield Council on Bioethics points to *"the faltering ability of conventional information governance measures to keep pace*

P. Vassilakopoulou (✉) • E. Skorve • M. Aanestad
Department of Informatics, University of Oslo, Postboks 1080, Oslo 0316, Norway
e-mail: xvasil@ifi.uio.no; espesko@ifi.uio.no; margunn@ifi.uio.no

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_11

with these developments" as a significant problem (Nuffield Council on Bioethics 2015, p.xvi). In this chapter we examine some of the emerging tensions that relate to genetic data ownership and sharing. Our case concerns information repositories for genetic mutation data related primarily to two specific genes which influence humans' susceptibility to breast and ovarian cancer (*BRCA1* and *BRCA2*). The case description spans the two decades that have elapsed since the two genes' identification. During this period, scientists around the world have been analysing the genetic material of thousands of individuals contributing to the accumulation of rich datasets, valuable for better understanding cancer biology and for diagnosing cancer susceptibility. The governance of these rich data sets has been the subject of controversies, contestations and bold initiatives from the onset of BRCA genetic testing till this day. Data governance arrangements impact the processes through which data acquire value, the distribution of this value, and the dynamics of value accrual. The longitudinal study on the evolution of BRCA data sharing is not only interesting for historical reasons but more importantly, it is a precondition for developing an understanding of the dynamics in the field, the different value logics and the tensions between them. This understanding is needed for developing and discussing premises for governing the collection, sharing and re-use of genetic datasets.

The chapter is structured as follows: in Sect. 2 we present the evolution of different BRCA dataset repositories and related data governance experiences, in Sect. 3 we analyse these experiences, elaborate on the unique characteristics of BRCA data and identify different logics shaping the field, then, in Sect. 4 we suggest drawing from the literature on collective action and the governance of commons for addressing the multiplicity of logics identified. We conclude by discussing the usefulness of foregrounding fundamental questions related to equity, efficiency and sustainability for shaping and evaluating governance arrangements in the field.

2 BRCA Mutation Data Generation and Sharing

2.1 The Advent of BRCA Testing

In 1994, after an intensive race among scientific laboratories around the world, a gene which influences susceptibility to breast and ovarian cancer was identified: *BRCA1* (Miki et al. 1994). Soon after that, a second gene related to susceptibility to the same diseases (*BRCA2*) was also identified (Wooster et al. 1995). The end of lengthy and contentious efforts within research labs marked the beginning of a new era for genetic testing. Tests could be extended beyond diseases that are inherited from generation to generation through simple patterns (like thalassemia or cystic fibrosis) to diseases where inheritance and environment are dual influences. Consequently, genetic testing could be relevant for a much wider population than the one affected by rare inheritable diseases. Announcing their identification of

the *BRCA1* gene, scientists noted that although their finding advances medical and biological knowledge, *"it also raises numerous ethical and practical issues, both scientific and social, that must be addressed by the medical community"* (Miki et al. 1994, p. 71).

Genetic testing in a clinical context entails mapping the sequence of the genes for a specific individual and comparing it with what is most commonly encountered in the general population. Differentiations (variants) from the common sequence are assessed by experts and classified as: variants that indicate pathogenicity; variants that do not indicate pathogenicity; or variants of uncertain/unclassified clinical significance (VUS). Based on genetic test results, doctors can plan clinical actions for different individuals; in the case of BRCA testing some of these actions can be as radical as mastectomy or ovary removal. Variant assessment entails the combination of information from multiple data sources including prior assessments performed in laboratories around the world, access to prior assessments can facilitate and accelerate the process significantly.

2.2 The Creation of a Common BRCA Data Repository: BIC

In 1995 a central, web-accessible repository named Breast Cancer Information Core (BIC) was created for hosting BRCA variant data. This was a grassroots initiative of ten scientists from different institutions (universities, hospitals, a research institute and a private company) in different countries. Scientists were motivated to create the novel common repository because the rapid rate of findings and the establishment of teams all over the world have rendered "*traditional sources of information — such as books and journals inadequate"* (Friend et al. 1995). A paper in *Human Mutation* from 1996 points not only to BIC's role for rapid information dissemination but also to its role for work coordination across sites: "*One of the serious impediments to achieving clinical benefits from the isolation of the BRCAl gene is finding and assessing the significance of mutations in this new cancer susceptibility gene. This will be greatly facilitated by coordinated detection and interpretation of mutations and the dissemination of this information to as many qualified investigators as possible. To this end, the BIC has created and maintains a central repository for information regarding mutations and polymorphism*" (Couch and Weber 1996). BIC hosts data deposited by individual investigators, research, hospital-based and commercial labs. The information content of each BIC entry comes out of scientific production processes actualised in different laboratories around the world and reflects the capabilities of these laboratories. The overall quality and utility of the repository relies to the steady inflow of up-to-date data from tests performed around the world. According to BIC guidelines, before data are added, they are examined and edited by several members of its steering committee. Registration is open to all. Registered users' access is unrestricted, but there are guidelines for data acknowledgement and use.

The creation of BIC made possible online information sharing within a community of researchers from all over the world. BIC contains today (2015) more than 30 000 entries with information on *BRCA1* and *BRCA2* variants, characterisations of variants' clinical significance, complementary data about samples' origin, detection methods, and depositor's contact details. This information has been contributed by scientists from more than 30 countries. Issues related to data registration harmonisation and completeness do exist, but, the main issue with the content of BIC relates to the high percentage of variants of "unknown clinical significance". Currently, around 30 % of BIC entries (corresponding to around 50 % of BIC variants) fall in this category. Practically, the information conveyed for variants of "unknown clinical significance" is (a) that they have been detected and (b) that the evidence collected had not been sufficient for their assessment. The information content of BIC in these cases does little to reduce the uncertainty in assessing the variants for cancer predisposition.

2.3 BIC's Main Depositor Pulling Out

Myriad Genetics which is the largest molecular diagnostic clinical laboratory in the world (Myriad Genetics 2014) was by far the largest data depositor to BIC up to year 2004 when they stopped contributing. Myriad Genetics was founded in 1991 as a spin-off from the University of Utah's Centre for Genetic Epidemiology. During the 1997–2000 period the company was granted nine patents in the United States on the *BRCA1* and *BRCA2* genes which gave control over the use of diagnostic tests on those genes (Gold and Carbone 2010). Myriad started offering genetic tests for clinical purposes through physicians and became the biggest contributor to the BIC database. Although the company did not prevent other laboratories from performing BRCA tests for research purposes they used their patents to control testing for clinical purposes (Gold and Carbone 2010) practically excluding other private companies from this domain. Myriad's dominant position in the USA explains its large number of submissions to BIC: more than 50 % of BIC entries during 1997–2004 came from Myriad. Obviously, being part of a community where all others could contribute significantly less was not a strong incentive to continue contributing. One of the members of BIC's steering committee puts it blandly: "*each laboratory has an incentive to do so in order to benefit from the data contributed by others and to further scientific progress. This incentive evaporates when one laboratory controls most of the data*" (Swisher 2009). Since late 2004 Myriad has not deposited data into BIC.

Myriad explained its decision to stop contributing data to BIC by claiming that the common database lacked operational and clinical standards, was subject to funding cuts, and could not handle the volume of data contributed by the company (Tucker 2014). The company also pointed to the high rate of variants classified as of unknown clinical significance in the public databases (Angrist and Cook-Deegan 2014). Myriad has published that from 2002 to 2013, they achieved an 84 % decline

in the rate of variants of uncertain significance in their test results (Eggington et al. 2013). For 2013, the rate of BRCA variants of uncertain significance in the tests performed by the company was down to 2.1 % (idem). To minimise uncertainty Myriad utilises multiple lines of evidence to evaluate and reclassify variants such as: in vitro assessments, segregation analysis, co-evaluation with mutations in other genes associated with the same syndrome, and the application of a scoring algorithm developed using the large numbers of data available to the company (Eggington et al. 2012, 2013; Pruss et al. 2014).

In 2013, the US Supreme Court invalidated Myriad's patents on *BRCA1* and *BRCA2* genes. A number of labs announced BRCA testing after the decision (Bravin and Kendall 2013). Nevertheless, Myriad has already a competitive advantage in BRCA testing by accumulating years of experience and building a proprietary database: "*by becoming the world's largest testing service, Myriad also discovers new variants and incorporates those into its database . . . For other genetic conditions, clinical interpretation is largely based on public data; for BRCA testing, Myriad has a distinct advantage, even over the best academic centers, because of its unique data set*" (Baldwin and Cook-Deegan 2013).

2.4 New Data Inflows: Reaching Out to Genetic Test Recipients

As a reaction to Myriad's decision to stop sharing information on BRCA variants, two senior medical geneticists initiated a project to collect reports from BRCA tests performed by the company in an alternative way. The project was named "Sharing Clinical Reports Project" (SCRP) and was addressed to physicians that receive test reports from Myriad (Nguyen and Terry 2013). The idea for this initiative is that even though data are not deposited directly to any common repository by the company they can still be collected via the distributed network of physicians that keep records of individual reports for their patients. Physicians that participate in SCRP are given two options: they can either forward full reports received from Myriad after stripping them off from patients' identification information or they can submit summary data extracted from the reports in a structured format (Sharing Clinical Reports Project 2014a). SCRP was initiated in 2012 (Kolata 2013) and up to August 2014 the project received 5416 submissions related to 2048 unique BRCA variants (Sharing Clinical Reports Project 2014b). Thirty four percent of these variants were assessed by Myriad as of "uncertain significance" (idem). Additionally, SCRP initiators started a "sister project" addressed to patients which was named: "Free the Data" (Free the Data 2014). This additional initiative has a great potential because as of February 2014, the US Department of Health and Human Services passed a new rule which provides for patient access to their test reports directly from the laboratory that performed it (before that, physicians were information gatekeepers) (Angrist and Cook-Deegan 2014). The "Free the Data" project not only provides a platform for patients to upload their BRCA reports but goes one step further by giving them the option to contribute additional personal health related information.

2.5 New Regimes for Common Genetic Data Repositories: ClinVar and BRCA Share

In 2012, the ClinGen initiative by the US National Center for Biotechnology Information (NCBI) released the prototype of a new open repository for data on variant interpretations (Landrum et al. 2013). After a brief piloting period it was officially launched in April 2013. This new repository (named ClinVar) invited scientific groups that hold data from clinical testing, research and literature to submit variants' interpretations. ClinVar is a variant-centric repository that covers all human gene variations – it is not dedicated to the BRCA genes. As of November 2014, it contained information on 125,588 variants of 19,753 different genes collected through 144,326 submissions (ClinVar 2014). Both BIC and SCRP registered as submitters and have contributed a significant volume of data on BRCA variants (approximately 5900 submissions as of November 2014). Unlike BIC where the content of submissions is reviewed by members of its steering committee and the variant interpretation for each entry is periodically updated to reflect the latest findings, in ClinVar the original content of the submissions is not curated or modified but only flagged with a review status. The status is practically indicating if the variant interpretation can be attributed to a single submitter, multiple ones or to a cross-institutional panel and if there are conflicting interpretations. Available interpretations for a particular variant are presented in comprehensive views where multiple submissions are associated. ClinVar supports the surfacing of conflicting interpretations and their eventual resolution by the scientific community on a case by case basis without the engagement of knowledgeable mediators since ClinVar does not arbitrate and does not resolve conflicts. Resolutions are provided by expert panels or professional societies (Landrum et al. 2013).

This new repository is practically a data consolidator that does not filter or control the data accumulated. Sceptics express concerns about overreliance to submitters and raise questions about the appropriateness of the free public access model. A medical genetics' specialist expressed his concerns in a paper published in the *Nature* journal soon after ClinVar's release: "*There is no revenue stream to pay an expert to review the data because you can get the data for free in ClinVar (. . .) this could paradoxically be a way in which the interpretation of variants ceases*" (Baker 2012). Despite the challenges, ClinVar quickly established a role in medical geneticists' everyday work. In a recent survey among medical genetics scientists working for research and healthcare, ClinVar was found to be the third most accessed resource for human genomic data (accessed by 33 % of survey respondents) (van Schaik et al. 2014).

An alternative model for genetic data accumulation and sharing is being employed by another recent initiative which was announced in April 2015. The commercial laboratory Quest Diagnostics and Inserm (the French National Institute of Health and Medical Research institution) announced the launch of BRCA Share. This is a public-private data sharing initiative which aims to provide scientists and laboratory organisations around the world with access to *BRCA1* and *BRCA2*

genetic data, with Quest licensing the data and forming sub-license agreements with participants. Content-wise, it builds upon the Universal Mutation Database that has been built through data sharing between 16 French laboratories over a decade, based on Inserm's gene-data curation process. To participate in BRCA Share, participants will commit to sharing past, present and future data in order to get access to the database. Commercial labs will pay according to their size, while research entities will get access at no cost. This financial arrangement allows BRCA Share to emphasize data curation arrangements to attend to data quality and to conduct functional studies on the effects of mutations, without depending on research or public funding. This alternative model was not received positively by all: a *Nature* editorial describes the BRCA Share initiative as a "walled garden" and laments its lack of sharing with the ClinVar open database (Nature Editorial 2015).

Throughout the past two decades various alternative configurations for BRCA data collection and sharing have been established. In each of those configurations the distribution of rights and obligations for data depositing, accessing and curating differs significantly. In the table that follows (Table 1) we summarise these key aspects for each configuration.

3 Unique Data Characteristics and Diverse Actors' Logics

3.1 Data Characteristics

The emergence of BRCA data repositories over the past two decades is in itself a clear indication that access to such data is considered a valuable resource amongst the various stakeholders involved. However, what constitutes the value of such access can vary across stakeholder groups. For clinical purposes, the reuse of previous assessments of clinical significance is core. For research purposes, the fact that a variant has – or has not – been previously observed can be of value in its own right. BRCA data sharing means that variant-specific information can be looked-up by scientists when a specific variant is re-encountered during testing. Scientists evaluate past variant assessments and use available data as resources to facilitate and enhance their own assessment. Due to the nature of the BRCA genes there are thousands of possible variants, and new rare variants are continuously identified. Both *BRCA1* and *BRCA2* are relatively "large-sized" genes (*BRCA1* has 24 exons, *BRCA2* has 27 exons) that include unusually long exons 11 where high rates of harmful variation occur. Examples of variations are: changes in sequence, changes in amount, and changes in position (Human Genome Variation Society 2007). Looking-up variants in the common repositories is possible because they are expressed according to one or more standard typologies in structured ways. This makes possible the unequivocal sharing of data on such complex scientific objects (although the relevant nomenclatures are still being extended and revised). *BRCA1* and *BRCA2* are related to monogenic predisposition syndromes (e.g. for

Table 1 Alternative configurations for BRCA data collection and sharing

	Description	Depositing	Accessing and reuse	Curating
Breast Information Core (BIC)	Repository for data on BRCA variant occurrences and their assessments.	Any genetic test performer after a registration process (individual investigators, research, clinical and commercial labs).	Anyone after a registration process. Unrestricted reuse if source is acknowledged.	By members on a voluntary basis.
Myriad's database	Repository for data on BRCA variant occurrences and their assessments.	One specific branch of private genetic testing labs.	Within the company.	Within the company.
Sharing Clinical Reports (SCRP) and Free the Data	Initiative for the collection of data on BRCA variant occurrences and their assessments.	Any genetic test recipient (clinicians and patients).	Unrestricted (data are submitted to ClinVar).	No curation.
ClinVar	Repository for consolidated data on variant assessments.	Any scientific group that holds assessment data (from clinical testing, research, literature).	Unrestricted.	No curation.
BRCA share	Repository for data on BRCA variant occurrences and their assessments	Genetic test performers (research and clinical, private and public) that commit to share past, present and future data.	Only for participants: paid access for commercial labs, no cost for research entities.	Professionally curated – curation is part of the configuration.

breast cancer), meaning that a positive find in this gene can be conclusive without further analysis of the genome. Practically, this means that the assessment of a variant's clinical significance should be valid for any person anywhere, i.e. variant assessments are globally relevant.

To sum up, BRCA variant assessment data have a number of interesting characteristics: their significance is high because they convey information related to human life and disease susceptibility, they are products of highly specialised scientific knowledge production processes, they relate to specific genetic material that shows high variability resulting to high numbers of rare occurrences, they convey information which is globally relevant and they are portable and combinable due to

standardisation. Practically, each new variant assessment is both a person-specific diagnosis and a contribution to general body of genetic knowledge. This is one of the reasons why BRCA data are objectified and sought after. By conceptualising the information content of BRCA data is in this way, the moral obligation to make this information available to anyone that needs to decide on a medical course of action after genetic testing can be argued. This information is not simply an outcome of routine lab activities but entails the exercise of scientific judgement, the employment of specialised support systems and reflects the capabilities and resources available in the lab where the data are generated. Labs that aim to maintain a competitive edge (ensuring their subsistence) are reluctant to share their data and make them available for reuse.

3.2 Research, Commercial and Clinical Logics at Play

The various groups of actors in the field operate according to different interests and agendas, and this shapes their stance in relation to data sharing. We will in the following distinguish between three different logics; a research logic, a commercial logic and a clinical logic. These logics differ, overlap, and engage with each other in non-trivial ways.

A *research logic* has the primary aim to further knowledge production, and thus openness, sharing and transparency are highly valued. The orientation towards data openness in genetic research is exemplified by the Bermuda principles that were established in 1996 within the Human Genome Project and stipulated that DNA sequence data should be published (i.e. uploaded to public repositories) within 24 h of production. These principles introduced institutional measures for rapid, pre-publication data sharing, not only post-publication that had been common (Contreras 2011) and were adopted by funding and policy entities such as e.g. the US National Institute of Health and the NHGRI (National Human Genome Research Institute). In the research field, funders and publishers of research are in a position of power from which they can enforce data sharing. Scientific journals demand disclosure upon publication. Human Mutation was the first journal to adopt a full data sharing requirement, in 2010, and the European Journal of Human Genetics has gone a step further, hiring curators to check that each paper's variant descriptions have been accurately transmitted to a public database (Krol 2014). This regulatory model which ensures data sharing through demands to deposit data in public repositories before a paper is accepted for publishing brings with it some risks. The primary motivation for data sharing may be to get the paper published, and the need to carefully curate, maintain and update the data deposited may take a secondary importance. This, together with lack of sustained and predictable funding, may explain the quality concerns of such open databases.

Within a *commercial logic* proprietary access to data equals competitive advantage and profit potential, as the case of Myriad exemplifies. Furthermore, a commercial company must have sufficient public trust and confidence. Myriad

emphasizes the quality and reliability of the information contained in its own variant databases to gain customers' trust. For instance, in a recent study conducted by Myriad, the quality of data is compared across five different BRCA related databases, BIC, ClinVar, HGMD Pro, LOVD and UMD (Vail et al. 2015). The researchers investigated a set of 1.327 variants from Myriad's database comparing the classifications for internal and cross-database consistency. The general tendency was that different sources more often than not reach what the authors characterise as conflicting conclusions. Based on these findings Vail et al. argue that lack of governance mechanisms that can assure proper quality control and adherence to current standards regarding methods, nomenclatures and documentation, render the open databases a potentially dangerous source of information for clinical purposes and that "healthcare providers should exercise caution when using these research tools for clinical purposes."

A *clinical logic* primarily seeks resolution of questions regarding an individual patient's situation. The work of a healthcare provider is characterised by a pragmatic search for relevant input to decision making following the Hippocratic mind-set of benefitting and never harming the patient. This includes seeking for evidence in the genetic data repositories. Given the varying quality of the available information, a clinician that approaches the market to order a genetic test for a patient is thus susceptible to the 'superior quality' argument that a specialised company like Myriad can offer. Moreover, within this clinical logic, the primary aim is not to contribute to knowledge production. This goes for individual clinicians that use commercial services as well as for clinical laboratories that do in-house testing, e.g. clinical genetic departments in hospitals. In these laboratories there may be researchers who publish scientific results, but there are no comprehensive mechanisms to ensure that the knowledge accumulated via the routine work of clinical testing of patients is shared. Usually scientists within clinical laboratories do look-up the variants identified in the open genetic repositories (so they have a strong interest in having access to such repositories) but their own data including their assessments of variants' significance are accumulated in local tools (e.g. in-house databases). For non-researchers there are few, if any, incentives to deposit these data in shared repositories. The ensuing "silo effect" of isolated repositories is problematic when we deal with a body of knowledge that is still young and under rapid and continuous development and growth claims (Rehm et al. 2015) as it can have life-or-death consequences for patients.

We have singled-out the three more influential logics shaping the field (research logic, commercial logic and clinical logic). Nevertheless, in the current context of clinical utilisation of genetic data there are more actors that add complexity to the picture. The perspectives of individual patients, public health policy makers, regulatory agencies, and funding bodies (public and private insurance funds) are also important. The multiplicity of actors with different resources, interests, engagement modes and time-horizons combined with the exceptional connectivity afforded by current information and communication technologies and the continuous advancements in gene sequencing technologies create a very dynamic and complex landscape.

4 Grappling with Different Logics

The different logics stem from actors' different positions and roles and together with the specific characteristics of the BRCA genetic data they shape the different repository configurations. There is a however a common interest in a continuously evolving body of knowledge. For sustaining and cultivating this body of knowledge there is a need for coordinating mechanisms. Instead of letting the logics play out unorderly, an overall approach for governing the field of collection and sharing of genetic data is needed. Hence, guiding principles for the development of governance arrangements have to be identified.

4.1 The Shortcomings of Existing Governance

Data sharing that ensures access to data for all is more or less enforced through funders' or publishers' policies. However, data are generated not only from research studies but also from clinical testing. To counter the proprietary incentives of for-profit actors who gather data from clinical testing of patients, Cook-Deegan et al. (2013) propose that payers (national health systems and insurers) and regulators should require data sharing from clinical testing as a condition for reimbursement, payment or allowing market access. Matloff et al. (2014) published in one of the domain journals an "ethical call for action" aimed to payers but also to referring clinicians (who choose laboratories for testing) urging them to make "the choice to use only laboratories that are committed to quality, efficiency, and facilitating progress for all through sharing of data". In February 2014, the Cancer Genetic Counselling Program of Yale School of Medicine issued a position statement where it is declared that laboratory choices will be made on the basis of four criteria: quality, time, cost and open access: "whenever possible we will choose laboratories that have pledged to make all of their past, present, and future gene data publicly available in order to allow this important information to be freely accessible to all clinicians and researchers, to further the advancement of medical knowledge and to best serve patient care. We will not support laboratories that hoard data" (Yale Cancer Genetic Counceling 2014).

The common feature of these initiatives is their aim to establish principles for the relationship between clinical and commercial actors that will make possible the accumulation of data in open repositories. They do not address research actors since data sharing has been generally recognised as important in medical genetics research, not only for the advancement of scientific knowledge, but also for the preservation of information and for safeguarding against misconduct (van Schaik et al. 2014). As discussed in Sect. 3.2 there are already a number of measures in place within the research domain seeking to enforce contribution to the shared resource, however, there is a lack of mechanisms to ensure continued attention to data after publication. This lack of maintenance undermines data quality and

consequently hampers the sustainability of the resource. Other measures, such as curation and quality assessment are needed to counter the tendency towards degradation of the shared repositories. This requires resources to be allocated specifically to curation and continuous improvement initiatives, whether from public or other type of pooled funding, volunteering (as in BIC) or from membership fees (as e.g. with BRCA Share). Sustainability and continued relevance of shared information resources require the ability of the resource to update the content (e.g. a mechanism for assessing and revising classifications of mutations) and to accommodate novel needs (e.g. complement data with outcomes of functional analyses). ClinVar's listing of the various (and sometimes contradictory) classifications for a single mutation is a way to open up for shared assessment and allows the shared knowledge to be continuously revised and amended.

4.2 Genetic Data Repositories as Common Good Resources

The previous section revealed some of the on-going struggles related to the establishment of principles and rules that can be used as a shared platform in the appropriation of effective governance principles. In order to understand the character of these struggles and why they seem so hard to reconcile, we now take a closer look at the very nature of genetic data and their special characteristics and discuss how the data repositories can be seen as a common good resource. These data are products of highly specialised scientific knowledge production processes that are possible because of the human and social capital that was devoted in the past to the advancement of methods and tools for human material analysing and interpreting. Their significance is high because they convey information related to human life and disease susceptibility. They are globally relevant, portable and combinable. They are sought after by multiple different actors that have diverse motivations, concerns and pressures but a common interest on sustaining and further developing this body of knowledge.

Whenever multiple actors build upon the same resource, potential problems exist (Hess and Ostrom 2006a): energy and work must be devoted to producing and effectively managing it, there can be incentives to free ride on the production process or to carelessly generate pollution (data that convey misinformation or incomplete data). Nevertheless, there is a common interest on this resource (the body of knowledge), it has significant societal value and fragmentation hampers the realization of its full potential, especially when some fragments are subjected to enclosure. Hence it makes sense to discuss this body of knowledge as a common good, susceptible to the same governance considerations as other common good resources. This entails some form of collective action, which in turn requires rules on issues of ownership, participation and responsibility. Based on this, we suggest drawing from the literature on collective action and the governance of commons

to gather intellectual resources for thinking about appropriate governance also in this field. This literature grapples "*with the age-old problem of how to induce collaborative problem solving and other forms of collective action among self-interested individuals, groups, or organizations, assuming, of course, that they share at least some common goals*" (Fulk et al. 1996).

Ellinor Ostrom's identification of what characterises robust, long-enduring, common-pool resource institutions led to the formulation of design principles for governance of such resources (Ostrom 1990). While traditional studies of commons (including Ostrom's) addressed natural resources, such as fisheries, fresh water and grazing land, also, information and knowledge – also within bioinformatics – have been studied from a commons or public goods perspective (see e.g. Fulk et al. 1996; Hess and Ostrom 2003, 2006a, b; Wasko and Faraj 2000). Indeed, current attempts at privatising knowledge resources has even been characterised a second wave of enclosure, with the fencing in of common land that started in the fifteenth century constituting the first wave (Boyle 2003). While there are clear differences between knowledge and natural resources, there are also significant similarities when such resources – or their utilisation – are of general public interest.

Previous research has pointed out the political (Winner 1980) and value laden (Van den Hoven 2007) aspects of information and communication technologies (ICTs). Rather than systems of improvement, the design and use of such technologies can be considered systems of re-distribution (Vikkelsø 2005). The regulatory power of these technologies (Lessig 1999, 2006) – combined with their tendency to embed and promote the world view and values of their creators (Van den Hoven 2007) – makes their entire lifecycle a subject to ethical considerations and potential dilemmas. This is no less true for the global knowledge infrastructure of our study than it is for the ICTs the above arguments are based on. While ICTs certainly constitutes part of this infrastructure, equally important are other elements, such as institutions, organisations, standards, procedures and other problem solving arrangements. From this emerges a need for responsible governance regimes that are capable of ensuring the overarching values this infrastructure is intended to promote. Hence, the question of what these values are becomes pivotal.

It is reasonable to assume a multitude of economic and professional values associated with the various actors' engagement with the infrastructure, but none of these can be considered *raison d'être* for the infrastructure as such. However, the scientific character of the knowledge draws attention to: (1) the value of scientific knowledge creation in its own right and (2) the application of scientific knowledge in efforts to improve human life. From this perspective, we argue that the over-arching purpose of the infrastructure is to facilitate the pursuit and application of scientific knowledge, and that its value subsequently accrues from its ability to do so. The more a value refers to a common good, the stronger it becomes as a justifying logic (Bergquist et al. 2012), hence this is the level of abstraction where we can find non-contested values upon which a governance regime can possibly achieve the coordination mechanisms needed for a less fragmented body of knowledge.

4.3 Equity, Efficiency and Sustainability

The commons literature points to equity, efficiency and sustainability as the fundamental issues that must be addressed by any common good governance regime (Hess and Ostrom 2006c).

Equity refers to the distribution of rights and obligations related to contributions, maintenance and use of a common good. Who does the work and who reaps the benefits? Is the resource readily available when needed for all members of the community? Are all voices heard when significant decisions are made regarding the governance of the common good? These are all important issues, not just to ensure democratic values as a founding principle, but also because they address the preconditions for commitment in an environment where few sanctions are available to force adherence to the community. The community in this regard will more often than not be a community without unity (Corlett 1989), and interdependence will then be what ties the community together (Kearney and Berkes 2007). While Myriad doesn't depend on others to provide information about genetic variants, the knowledge it builds in-house is still based on the current advances in the field as such. And though it has been able to use its market domination in the area of BRCA testing to keep parts of its knowledge enclosed, there is still a request for legitimation of this choice coming from the wider community. Indeed, what we see now is attempts from the community to throttle access to resources such as reimbursements and scientific publication outlets, in order to instigate a change in practices based on disclosure policies. In contrast to Myriad's approach, the ClinVar repository is fully open access, anyone can register with BIC in order to get access rights, and BRCA Share is open to anyone committing to contribute their own current and future knowledge regarding BRCA mutations. While the two latter might resemble a 'clubbing' of the knowledge commons (Hess and Ostrom 2006c), the relative ease of gaining access justifies regarding these repositories as part of the common goods. *Efficiency* refers to a governance regime's ability to ensure an optimal management of the common good. Optimal in this context does not mean perfect, but rather a pragmatic approach to a solution close to what is possible to achieve under current conditions. Does the current arrangement work? Are the benefits proportional to the costs and efforts – locally and globally? And especially relevant in the context of knowledge resources; are we able to realise the potential synergies offered by the diversity of our community? Does the quality and value of the resources meet the requirements of their users? The latter of these questions can be especially tricky to answer, as what constitutes quality and value will vary across groups and logics. For instance, while uncertainty related to clinical conclusions pose potential research questions rather than problems within a research logic, such uncertainty can be very problematic within a clinical logic (as reflected in Vail et al.'s (2015) warning against relying on open access databases for clinical purposes).

Sustainability refers to the governance regime's ability to ensure a responsible resource trajectory in a long time perspective. While resource depletion is not

an issue when knowledge is the resource, the field of genetic analysis is still young and the community depends on continuous advancements in order to meet future needs and obligations. As knowledge is a resource primarily developed based on what is already known, any enclosure policy is thus clearly an obstacle to a sustainable development. But equally important is the current arrangement's ability to support aggregation of existing knowledge as a foundation for knowledge creation. A crucial aspect relates to the ability of the information resources to continue to grow, maintain a high quality and be able to accommodate novel data needs. For this, attention to interoperability, data quality and diversity of knowledge providers are core issues. The arrival of new actors and interests emerging from e.g. multi-gene panel testing, whole-genome and whole-exome testing are examples of issues likely to emerge in the near future. The governance regime's ability to accommodate these will have an impact on the trajectory of the knowledge commons. At a meta-level, trust and commitment are also significant elements in assuring a sustainable development, thus tying sustainability tightly to the principles of equity and efficiency. Also, in complex systems diversity generates resilience, and resilience is a precondition for sustainability (Brand 1999). A governance regime that grants some actors a dominant position can jeopardise this diversity, possibly at the cost of a sustainable trajectory. From this perspective, the multiple logics and complexity of the system that might pose difficulties related to equity and efficiency, becomes a potential asset that can ensure sustainability, but only if all voices are heard and complexity is reflected in the governance models (Skorve 2013).

5 Concluding Discussion and a Way Forward

The narrative we have presented is illustrative of the growing complexity in the creation, maintenance and dissemination of genetic knowledge. Our longitudinal inquiry into the evolution of BRCA genetic data repositories reveals the concurrency of different logics. This is a field where it is difficult to find a common basis for resolutions since multiple actors' intentions are simultaneously pursued. Trying to reconcile diverse pursuits like clinical results, knowledge generation, commercial profit but also, cost containment for payers, patients' trust and society's confidence is not a trivial matter.

With the current advent of next generation sequencing, and the ever increasing areas where genetic testing is considered relevant for clinical purposes, it is not unreasonable to expect this complexity to grow exponentially. As for most complex systems, we have no doubt that this one will eventually find ways to adapt to the rapid changes it takes part in creating. The question is what form and how long this will take.

It is our contention that a consciousness regarding these issues from an early outset can both ease the transition into an adaptive system, and also drive this development towards arrangements where ethical considerations become an embedded

aspect of decision making. By discussing the global body of genetic knowledge as a common good, subject to the same governance considerations as other common good resources, we hope to contribute to this. The fundamental questions related to equity, efficiency and sustainability are, we believe, useful for research and practice alike when shaping and evaluating the current and future trajectories for this common good.

Acknowledgement We gratefully acknowledge Morten Christoph Eike for fruitful discussions and for commenting on this chapter, and Sarah Louise Ariansen, Sheba Maria Lothe, Thomas B. Grünfeld, Dag Undlien, Tony Håndstad, Timothy Hughes, and Eidi Nafstad for sharing their domain expertise.

References

Angrist, M., and R. Cook-Deegan. 2014. Distributing the future: The weak justifications for keeping human genomic databases secret and the challenges and opportunities in reverse engineering them. *Applied & Translational Genomics* 3: 124–127.

Baker, M. 2012. One-stop shop for disease genes. *Nature* 491: 171.

Baldwin, A.L., and R. Cook-Deegan. 2013. Constructing narratives of heroism and villainy: Case study of Myriad's BRACAnalysis® compared to Genentech's Herceptin®. *Genome Medicine* 5: 8.

Bergquist, M., J. Ljungberg, and B. Rolandsson. 2012. Justifying the value of open source. *ECIS 2012 Proceedings. Paper* 122.

Boyle, J. 2003. The second enclosure movement and the construction of the public domain. *Law and Contemporary Problems* 66: 33–74.

Brand, S. 1999. *Clock of the long now: Time and responsibility*. New York: Basic Books.

Bravin, J., and B. Kendall. 2013. Justices strike down gene patents. *The Wall Street Journal* 13: 2013.

Clinvar. 2014. ClinVar submissions. Available: http://www.ncbi.nlm.nih.gov/clinvar/submitters/. Accessed 13 Nov 2014.

Contreras, J.L. 2011. Bermuda's legacy: Policy, patents, and the design of the Genome Commons. *Minnesota Journal of Law and Science & Techology* 12: 61.

Cook-Deegan, R., J.M. Conley, J.P. Evans, and D. Vorhaus. 2013. The next controversy in genetic testing: Clinical data as trade secrets&quest. *European Journal of Human Genetics* 21: 585–588.

Corlett, W. 1989. *Community without unity*. Durham: Duke University Press.

Couch, F., and B.L. Weber. 1996. Mutations and polymorphisms in the familial early-onset breast cancer (*BRCA1*) gene. *Human Mutation* 8: 8–18.

Eggington, J., L. Burbidge, B. Roa, D. Pruss, K. Bowles, E. Rosenthal, L. Esterling, and R. Wenstrup. 2012. Current variant of uncertain significance rates in BRCA1/2 and Lynch Syndrome testing (MLH1, MSH2, MSH6, PMS2, EPCAM). ACMG Annual Clinical Genetics Meeting, Charlotte, 2012.

Eggington, J., K. Bowles, K. Moyes, S. Manley, L. Esterling, S. Sizemore, E. Rosenthal, A. Theisen, J. Saam, C. Arnell, D. Pruss, J. Bennett, L. Burbidge, B. Roa, and R. Wenstrup. 2013. A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. *Clinical Genetics* 86: 229–237.

Free the Data. 2014. Genetic information is more valuable when shared. Available: http://www.free-the-data.org. Accessed 15 Sept 2014.

Friend, S., A.L. Borresen, L. Brody, G. Casey, P. Devilee, S. Gayther, D. Goldgar, P. Murphy, B.L. Weber, and R. Wiseman. 1995. Breast cancer information on the web. *Nature Genetics* 11: 238–239.

Fulk, J., A.J. Flanagin, M.E. Kalman, P.R. Monge, and T. Ryan. 1996. Connective and communal public goods in interactive communication systems. *Communication Theory* 6: 60–87.

Gold, E.R., and J. Carbone. 2010. Myriad genetics: In the eye of the policy storm. *Genetics in Medicine* 12: S39–S70.

Hess, C., and E. Ostrom. 2003. Ideas, artifacts, and facilities: Information as a common-pool resource. *Law and Contemporary Problems* 66: 111–145.

Hess, C., and E. Ostrom. 2006a. A framework for analysing the microbiological commons. *International Social Science Journal* 58: 335–349.

Hess, C., and E. Ostrom. 2006b. A framework for analyzing the knowledge commons. In *Understanding knowledge as a commons: From theory to practice*, ed. C. Hess and E. Ostrom. Cambridge, MA: MIT Press.

Hess, C., and E. Ostrom. 2006c. Introduction: An overview of the knowledge commons. In *Understanding knowledge as a commons: From theory to practice*, ed. C. Hess and E. Ostrom. Cambridge, MA: MIT Press.

Human Genome Variation Society. 2007. *Mutation nomenclature: Recommendations for the description of DNA changes*. Available: http://www.hgvs.org/mutnomen/ESHG2007_W12_JdD.pdf. Accessed 11 Sept 2014.

Kearney, J., and F. Berkes. 2007. Communities of interdependence for adaptive co-management. In *Adaptive co-management: Collaboration, learning and multi-level governance*, ed. D.B. Armitage, D. Armitage, and N. Doubleday. Vancouver: UBC Press.

Kolata, G. 2013. DNA project aims to make public a company's data on cancer genes. *The New York Times* [Online]. Available: http://www.nytimes.com/2013/04/13/health/dna-project-aims-to-make-companys-data-public.html?pagewanted=all&_r=0. Accessed 15 Sept 2014.

Krol, A. 2014. As genetics moves to the clinic, pathogenic variants still subject to doubt and debate. *Bio-IT World* [Online]. Available: http://www.bio-itworld.com/2014/4/17/genetics-moves-clinic-pathogenic-variants-still-subject-doubt-debate.html. Accessed 11 Nov 2014.

Landrum, M.J., J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, and D.R. Maglott. 2013. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* 42: D980–D985.

Lessig, L. 1999. *Code and other laws of cyberspace*. New York: Basic Books.

Lessig, L. 2006. *Code: Version 2.0*. New York: Basic Books.

Matloff, E., R. Barnett, and R. Nussbaum. 2014. Choosing a BRCA genetic testing laboratory: A patient-centric and ethical call to action for clinicians and payers. *Evidence-Based Oncology*. http://www.ajmc.com/journals/evidence-based-oncology/2014/may-2014/Choosing-a-BRCA-Genetic-Testing-Laboratory-A-Patient-Centric-and-Ethical-Call-to-Action-for-Clinicians-and-Payers. Accessed 23 May 2016.

Miki, Y., J. Swensen, D. Shattuck-Eidens, P.A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L.M. Bennett, and W. Ding. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* 266: 66–71.

Myriad Genetics. 2014. Myriad Genetics Corporate Presentation. https://myriad-web.s3.amazonaws.com/myriad.com/pdf/Myriad-Corporate-Presentation.pdf. Accessed 23 May 2016.

Nature Editorial. 2015. Thank you for sharing. *Nature* 520: 585.

Nguyen, S., and S. Terry. 2013. Free the data: The end of genetic data as trade secrets. *Genetic Testing and Molecular Biomarkers* 17: 579–580.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*. London: Nuffield Council on Bioethics.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York, NY: Cambridge University Press.

Pruss, D., B. Morris, E. Hughes, J. Eggington, L. Esterling, B. Robinson, A. Van Kan, P. Fernandes, B. Roa, and A. Gutin. 2014. Development and validation of a new algorithm for the reclassification of genetic variants identified in the *BRCA1* and *BRCA2* genes. *Breast Cancer Research and Treatment* 147: 119–132.

Rehm, H., J. Berg, L. Brooks, C. Bustamante, J.P. Evans, M.J. Landrum, D. Ledbetter, D.R. Maglott, C. Martin, and R. Nussbaum. 2015. ClinGen—the clinical genome resource. *New England Journal of Medicine* 372: 2235–2242.

Sharing Clinical Reports Project. 2014a. How to submit data. Available: http://www.iccg.org/about-the-iccg/collaborations/sharing-clinical-reports-project/how-to-submit-data/. Accessed 11 Sept 2014.

Sharing Clinical Reports Project. 2014b. Sharing clinical reports project. Available: http://www.iccg.org/about-the-iccg/collaborations/sharing-clinical-reports-project/. Accessed 15 Sept 2014.

Skorve, E. 2013. Shaping information infrastructures: Complexity and management in the context of a clinical IS implementation. PhD, University of Oslo, Oslo.

Swisher, E. 2009. Declaration in support to the Association for Molecular Pathology et al v. United States Patent and Trademark Office et al. United States District Court Southern District of New York.

Tucker, K.I. 2014. Genetics lab refuses to share data that could save lives. *Jewish Daily Forward* (12 August 2014).

Vail, P., B. Morris, A. Van Kan, B. Burdett, K. Moyes, A. Theisen, I.D. Kerr, R. Wenstrup, and J. Eggington. 2015. Comparison of locus-specific databases for *BRCA1* and *BRCA2* variants reveals disparity in variant classification within and among databases. *Journal of Community Genetics* 6: 1–9.

Van Den Hoven, J., and V. Laurent. 2007. ICT and value sensitive design. In *IFIP international federation for information processing, volume 233. The information society: Innovations, legitimacy, ethics and democracy*, ed. P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, and V. Laurent. Boston: Springer.

Van Schaik, T.A., N.V. Kovalevskaya, E. Protopapas, H. Wahid, and F.G.G. Nielsen. 2014. The need to redefine genomic data sharing: A focus on data accessibility. *Applied & Translational Genomics* 3: 100–104.

Vikkelsø, S. 2005. Subtle redistribution of work, attention and risks: Electronic patient records and organisational consequences. *Scandinavian Journal of Information Systems* 17: 10.

Wasko, M.M., and S. Faraj. 2000. "It is what one does": Why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems* 9: 155–173.

Winner, L. 1980. Do artifacts have politics? *Daedalus* 109: 121–136.

Wooster, R., G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem. 1995. Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* 378: 789–792.

Yale Cancer Genetic Counceling. 2014. Genetic Testing Lab Position Statement [Online]. Available: http://yalecancergeneticcounseling.blogspot.gr/2014/02/genetic-testing-lab-position-statement_2573.html. Accessed 7 July 2015.

State Responsibility and Accountability in Managing Big Data in Biobank Research: Tensions and Challenges in the Right of Access to Data

Aaro Tupasela and Sandra Liede

Abstract Within the European context the Data Protection Directive (Directive 95/46/EC) maintains an important role in current legal debates on the rights and obligations different stakeholders have in the processing of personal data. Biobanking and data sharing infrastructures pose new ethical and legal dilemmas in the interpretations we uphold with regard to the processing of personal data. This chapter examines the challenges associated with the data subject's right of access to data in Finnish biobanking. The Data Protection Directive provides provisions for individuals to confirm "as to whether or not data relating to him are being processed". Finland's recent Biobank Act (688/2012) has raised concerns since it also requires biobanks to provide, upon request, information regarding data which may have clinical (actionable) relevance for the individual's personal health. There is, however, no governance mechanism in place through which common standards and practices could be implemented. As a result the extension of the right to access data and mandating biobanks to relate significance of results to personal health has become a major concern for biobankers in Finland. The management of data, research results and incidental findings in biobanks is becoming, however, an increasingly significant challenge for all biobanks and the countries which are in the process of drafting policy and regulatory frameworks for the management and governance of big data, public health genomics and personalised medicine. The Finnish case highlights the challenges that many states are increasingly facing across Europe and elsewhere in terms of how to govern and coordinate the management of biomedical big data.

A. Tupasela (✉)
Department of Public Health, Centre for Medical Science and Technology Studies, University of Copenhagen, Øster Farimagsgade 5, Copenhagen K DK-1014, Denmark
e-mail: aatu@sund.ku.dk

S. Liede, LLM
Faculty of Law, University of Helsinki, Helsinki, Finland

Legal Affairs of Biobanking, National Supervisory Authority for Welfare and Health, Helsinki, Finland
e-mail: sandra.liede@valvira.fi

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_12

1 Introduction

During the past decade, ethical issues related to public health genomics[1] and biobanking have become increasingly prevalent in a number of countries around the world. This prevalence can be attributed to a number of factors including the increase in the number and size of biobanks, as well as the rapid development of analysis technologies, which now have made whole genome and exome sequencing practically and financially a viable option in research and diagnostics. For the Nordic countries especially, the questions are of major importance since individuals are given social security numbers at birth, and which are used to follow and track a broad range of data on them through-out their lives. This includes tax and income data, as well as healthcare data, which may also include tissue samples taken during routine diagnostics or when participating in research. Many of the debates and discussions flowing out of the post-genomic ethical landscape have sought to address issues such as sample and data use (Zika et al. 2010), secondary uses of samples (Cambon-Thomsen et al. 2007), as well as disclosure of genetic research results and incidental findings (Bledsoe et al. 2013; Wolfe et al. 2012; Knoppers et al. 2006). What is clear is that biomedicine is raising an increased amount of ethical issues related to the governance of big data (Mittelstadt and Floridi 2016). With regard to developing a public health genomics policy framework, Burke et al. (2010, 789) have suggested a genome-based agenda, which includes an infrastructure for generating an evidence-base for genomic medicine. Such big data policy frameworks are built around the notion of personalised medicine. In such future visions of public health genomics, personalised medicine has been seen to entail forms of empowerment among the patients, whereby they are seen to take increased responsibility for their health and lifestyle (Eric et al. 2012). This approach has also been seen as a form of neo-liberal health politics in which the individual are placed at the centre of decision-making.

Such big data policy agendas raise a more fundamental question for states with regard to the ways in which they envision the functioning of their healthcare systems in the future, as well as the relationship and infrastructure that will be constructed between research and treatment. It also raises a more general concern relating to the impact that size has on existing ethical and legal frameworks, as well as more mundane everyday research practices (Hoeyer 2012). On the one hand, individuals are expected to increasingly behave in more responsible ways, while the state still maintains an enormous and important role in the development and implementation of research and health care policy, particularly within the Nordic welfare state system. The management of big data is a challenging task since it requires large amounts of resources and infrastructure, and when implemented at the national level the whole population has the potential of becoming a cohort (Frank 2000). Numerous international organisations have sought to address the ways in which

[1]The responsible and effective translation of genome-based information and technologies for the benefit of population health (PACITA 2014).

healthcare services will be delivered in the future (European Commission 2013; EU Workshop 2003; PwC 2005). An OECD report (2013, 90), for example, has suggested that in relation to the development of personalised medicine in health care, the public needs to be educated with regard to genetic privacy protections. This would seem to suggest that new technologies bring with them a host of social, legal and ethical issues which need to be addressed at a broader societal level, and may not be manageable by experts and policy makers alone.

Concomitant to the explosion in the capacity to collect, generate and analyse genomic data has been the legal efforts to protect and safeguard the rights of individuals with respect to the collection, storage and processing of personal data. Emerging technologies, as well as politically heightened public health and big data issues function as incentives affecting legislators and regulators. However, a framework of European and International conventions, regulations and guidelines set the boundaries within which states and regulators are able to operate. The processing of human samples and personal data is regulated both at the local and International level by instruments that have various degrees of binding force.

Within the European context, the Council of Europe's (CoE) Biomedicine Convention (1997, ETS 164) and its additional protocols play a pivotal part in regulating the broad area of scientific research and advanced medicine as well as safeguarding decisional autonomy and protecting the rights of individuals, but it does not give specific guidance on how to implement the safeguards in practice at a national level. The Biomedicine Convention is built on four normative pillars, the right of protection for human dignity and identity, the right of respect for one´s integrity, the right to (equal access to) health care and the prohibition of non-discrimination (Dute 2005). The Convention and especially its Additional Protocol concerning biomedical research (Council of Europe 2005, ETS 195) cover all areas of biomedical research including use of personal data[2] collected for specific research projects and are effective in all the countries which have signed and ratified them.[3] Additionally, case law of the European Court of Human Rights (ECHR) relating to interpretation of the Biomedicine Convention extends the convention's influence to countries which have not signed or ratified it (ETS 164, Art. 29). As the scopes of application are limited to processing samples and data for specific research purposes, these legal instruments are not fully applicable in the big data sphere in which wide data sharing concepts are prevailing. However, the significance of the convention is emphasized in that it serves as a basis for a common approach to patients' rights and provides an International framework for health law as a legal discipline (Dute 2005). Furthermore, the principles of the Convention can serve as a guiding light for legislators in search for fair solutions for legal problems in healthcare systems that are in a transitional state.

[2]The Additional Protocol refers in art. 26 directly to law on the protection of individuals with regard to processing of personal data, whilst the Convention states respect for private life and right to information (art. 10).

[3]Finland has signed and ratified the Biomedicine Convention but not its additional protocol.

In the EU context, the Data Protection Directive (Directive 95/46/EC) maintains a leading role in current legal debates in safeguarding the collection, processing and sharing of various types of personal data. The draft Data Protection Regulation that is currently under consideration seeks a balance between the different cultures of managing personal data in the Nordic countries and the rest of Europe. The Nordic countries have had a long tradition of using personal social security numbers with which people can be connected and identified across numerous databases, ranging from tax and housing information to information related to health and medical treatment. The system of surveillance related to personal social security numbers is seen as a central feature of effective state administration in the Nordic countries and any attempt to make such systems less effective will be strongly contested by Nordic member states. These debates also draw heavily on notions of individual rights and autonomy. Under EU law, as well as under CoE law (1981, ETS 108),[4] 'personal data' is defined as information relating to an identified or identifiable natural person. Under the two branches of European law, there are categories of personal data which are considered sensitive by their nature and thus require special safeguards for ensuring the rights of the data subject. The Directive´s implications for processing data in biobanks and in other data sharing infrastructures, is understood by what it states on both general rules of processing non-sensitive 'normal' personal data (art. 7) and special rules of processing sensitive personal data (art. 8). The two-tiered classification indicates the impact processing could have on fundamental rights of a data subject (Hallinan and Friedewald 2015). The basis for legitimate data processing in both categories is specific consent, and in the case of art. 8, also explicit. Data processed by biobanks is mostly classified as sensitive, and hence, has to adhere to art. 8., which lays down the conditions for legitimate data processing. As Hallinan and Friedewald (2015) have argued, biobanking does not entirely meet the requirements of data protection regulation but are in the opinion that these requirements can be debatable. The aim of the Data Protection Directive has been to harmonize national data protection laws and to ensure equal level of protection in all Member States. However, the Directive has been considered outdated and does not sufficiently recognize rapid technological developments, globalization or biobanking. At the time of publication of this Chapter, a proposal for a General Data Protection Regulation is being drafted and is a policy priority for 2015 (cf. Hallinan and Friedewald 2015; http://ec.europa.eu/justice/data-protection/).

Biobanks have raised an important question as to what qualifies as data (e.g. a physical sample, data collected through analysis of the sample, genetic data, health and lifestyle data etc.), and whether different data types have implications as to the ways in which they are managed and governed (cf. Tupasela et al. 2010; Tupasela and Snell 2012). In Finland, a prominent argument made by some commentators has been that genetic and genomic data is the same as any other type of social or economic statistic which the state is able to collect and analyse (Aromaa et al. 2002).

[4]Especially the CoE Convention (ETS 108), which is the only legally binding international instrument in the data protection field.

This, however, contradicts with the wording of the draft General Data Protection Regulation, which considers genetic data sensitive and in need of special protection (Art. 9). Such conceptualizing of data into special categories can be criticized and a risk-based or knowledge-based approach to sensitive data processing could serve as a more rational way to adequately and equally protect the goals of future policies and regulations. The ambiguity as to the significance of the data that biobanks can produce can be seen as one of the contributing factors in trying to develop strategies to manage the data itself.

A number of countries have begun to develop and implement strategies through which genome research will be integrated into the healthcare system. In the UK, the 100 000 genomes project has started with the aim of sequencing 100 000 genomes which can be used for research (Genomics England 2015). Germany and the US have also established programs for supporting personalised healthcare by sequencing a large number of genomes to serve as both reference and variation databases. Likewise Finland has also recently published its own national genome strategy which will seek to build on its existing collections and population registers to provide better individual treatments (Ministry for Social Affairs and Health 2015a). What all of these projects and initiatives have in common is an ethical and legal pre-occupation with how to manage the vast amount of data that such research produces. More specifically, many such initiatives are setting the groundwork for what policies and strategies to adopt with relation to incidental findings (IF), return of individual research results (IRR), as well as interpretation of variants of uncertain significance (VUS). Although many of the strategies do not explicitly address such issues, there is an underlying assumption that new genomic data will open up the door for new forms of intervention.

In 2013 Finland was the first country to enact a comprehensive biobanking legislation, which covered a broad spectrum of tissue procurement, ranging from clinical to research samples for research purposes (Soini 2013; Tupasela 2015). The uniqueness of the legislation was in the fact that besides recognizing the right of individuals to enquire about the ways in which their samples were being used for research, it also allowed them to request that the biobank provide them with information regarding the significance of the findings to their health. Knoppers et al. (2006) have noted that, at the international level, there exists an ethical duty to return individual research results if the findings are valid, significant and there is proof that such findings were of benefit to the research subject or patient. In its wording, the Finnish legislation sought to recognize this duty, but placed the responsibility of action on the research participant or patient themselves, whereby they had to actively consent to, seek out and request such information. According to the legislation:

> A registered individual has the right to receive, upon request, information concerning his or her health as determined based on a sample. When providing information determined based on the sample, the person must be provided with an opportunity to receive an account of the significance of the information. A fee may be charged for clarifying the significance of the information that, at maximum, corresponds to the expenses incurred by providing the clarification (Biobank Act 688/2012, Section 39).

Additionally, an affiliated Government Decree (643/2013) states that a consent document shall include information of the sample donor's possible consent to receiving information on a clinically significant finding. The aforementioned Government Decree indicates an ethical assessment point for biobanks in that they may also themselves actively reach out to the registered person in order to report significant findings, if the person has consented to this beforehand and if the biobank evaluates the data to be relevant. However, the legislator has admitted in the preparatory work of the Biobank Act that there is a need to clarify to the data subject that biobanks do not have actual clinical ability to analyse the findings or their significance to a person.

This chapter interrogates the ethical dilemma and challenge in interpreting the European Data Directive with regard to biobanking in this manner and seeks to identify some of the pitfalls that it contains, both for patients and research subjects, but for biobanking as well. In addition, we seek to identify some of the policy challenges and problems that this approach entails with regard to the delivery of equitable healthcare services. We will first discuss the content and role of the new biobanking legislation in Finland with regard to the legal principles that it draws on. Following the discussion of the Finnish legislation we move on to a more general discussion on the governance challenges that states, organisations, as well as individual doctors may face when considering policies for the management of information derived from genomic analysis. Finally we make some conclusions and observations regarding the balance that we think are necessary with regard to interpreting personal privacy laws and the role of the state as a guarantor of equal access to healthcare services. We argue that with the ability to analyse and connect increasingly large amounts of data sets – such as with combining genomic data with register based data – there is a need for stronger state role in the coordination of various actors as to what principles and guidelines should be drawn upon in implementing broader directives on the management of personal data. This position also entails a clearer stance at the European level with regard to the new Data Directive in that biobanking operates in an increasingly international environment. States cannot, in our opinion, withdraw to a position of technical administration, but rather should play a stronger role in defining what principles ought to guide research activities and decision making on a more general level.

In our analysis we draw on what Hancher and Moran (1989) have called regulatory space to see how practical limits and thresholds of responsibility and accountability in reporting significant findings are exposed when left up to individual research subjects and patients. The regulatory space provide important insights into the limits of the welfare state system in delegating responsibility to individuals, as well as institutional actors, such as biobanks, when it comes to developing and setting up national infrastructures, which may have significant impacts on the ways individuals are managed within the healthcare system. The processes involved in sorting out roles and responsibilities in managing data also highlights the ways in which state authority still draws on broader principles of rights and duties when it comes to the general welfare of the population. The combination of these two perspectives and the long-term insights gained with working with

Finnish biobanking serves as the backdrop for our analysis into the challenges of implementing biobanking governance policies with regard to the management of personal data that is derived from research that draws on biobanked samples. In the following we will look at the challenges from a legal perspective to highlight the tensions which have arisen in the interpretation and implementation of the Biobanking Act in Finland.

2 Finland's New Biobanking Law

Since its adoption in September 2013, Finland's new biobanking law has been nationally hailed as 'the world's best act in biobanks' (Sitra 2014), which indicates how well the law has been welcomed by actors in the Finnish biobanking field. The Act aims to support research and promote openness during the use of samples, as well as secure individual rights such as privacy and self-determination. The Act covers all research activities ranging from basic to applied research and translational collaborative efforts resulting in new products and services (cf. Soini 2013). The Biobank Act can be characterized as an extension to Finland's existing research regulation giving researchers a wider opportunity to utilize the country's valuable sample and data materials.

Finland now has eight different types of high quality biobanks, ranging from large population based cohorts to hospital based biobanks and smaller disease specific biobanks. The growing infrastructure and operations are backed up by comprehensive regulation and supervision, which also function as a quality guarantee within the field. As the data being collected and generated by biobanking activities tend to mostly be sensitive in nature, principles of data protection have a strong influence within the field and the Biobank Act makes explicit reference to national data protection legislation. However, many of the provisions in the Biobank Act actually differ from general data processing prerequisites, which have clearly increased the need for guidance as to the interpretation of the Biobank Act in relation to data protection regulation. For example, biobanks have been granted leverage by enabling the use of wide consent instead of requiring specific and explicit consent for processing of sensitive data – which would be in line with data protection regulation (cf. Hallinan and Friedewald 2015).

By wide consent, the legislator has meant that samples and associated data can be collected for a wide range of future research purposes by using a general description of the intended research objective. The specification does not happen at an explicit research study level, but rather through a description of the research area which in the advantage of a biobank can be quite wide covering for example all research activities of a hospital district. The local data protection authority has noted that even though the given wide consent may legitimize processing of samples and associated data in a biobank, all other relevant data protection measures must be taken to ensure full protection of the data subjects. Security has been intensified with a double coding structure meaning not only have personal identifiers and samples

been separated into two different databases (for double coding: cf. Thorogood et al. 2014) but the samples are coded yet again once they are handed over to a research project – all to be reversed again by a keyholder for possible linkage. As wide consent has meant exemption from precise specification of sample and associated data types to be collected and utilized, authorities have been prudent in interpreting which type of data can be considered as *associated* with the samples (National Supervisory Authority for Welfare and Health 2015). As a counterweight to wide consent, and in addition to extensive supervision by health and data protection authorities, individuals in turn have been granted extensive control over use of their samples and data. This may be seen as a form of empowerment of autonomy but also as a shift in responsibilities between the individual and the state. The data subject has been granted legal power to monitor data processing, but at the same time an ethical responsibility to act on the information disclosed to him or her.

After nearly 2 years' experience of executing and interpreting the new biobanking legislation, it is relatively safe to say that the regulation has by no means been flawless. One of the major issues of conflict has risen from the individual's twofold right to request information from a biobank (right of access). *First*, according to the Biobank Act (Section 39.1), "everyone has a right to receive, upon request, information on whether or not his or her samples are being stored in a biobank, the legal grounds for storage, where associated data has been collected from and who it has been disclosed to." The provision does not extend to obligating a biobank to disclose specific details of a certain research project or research results.[5] The right to information is considered to be a key element of control and in realizing an individual's right to self-determination and may be executed through online access.

According to the preparatory documents (Governmental Proposal 86/2011) of the Biobank Act, Section 39.1 supplements the right of access as stated in Section 26 of the national Personal Data Act (523/1999). Section 26 is grounded in EU data protection regulation and states that "Regardless of secrecy provisions, everyone shall have the right of access, after having supplied sufficient search criteria, to *the data* on him/her in a personal data file, or to notice that the file contains no such data". The data subject shall be provided with information regarding regular data sources and destinations of disclosure. The wording of data protection regulation appears to apply widely to all data in a file, whereas the Biobank Act's scope of access is more limited. However, the Biobank Act has not intended to restrict the individual rights and autonomy granted in data protection legislation. Hence, there has been challenges in understanding the Finnish legislator's motivations for separately regulating the issue in the Biobank Act and how this right ought to be successfully interpreted either as an extension (supplement) of data protection regulation or as an independent provision.

Initially, actors within the biobanking field interpreted that the Biobank Act provided wider protection for individuals than the data protection regulation and

[5]However, publication of results and samples stored and used in a biobank is mandatory (Biobank Act, section 5) as the idea is to accumulate the data being generated in biobanks.

should *as lex specialis* be prioritized. The ultimate motivation for promoting this interpretation was unclear, but commercial interests may have influenced it, as the Biobank Act permits collecting a cost-corresponding fee for providing information, whereas the Personal Data Act provides the opportunity only once a year. The biobanking field placed special emphasis on the fact that data protection regulation enables exempting access for reasons stated in section 27 of the Personal Data Act. The Biobank Act doesn't include similar restrictions. Section 39.1 of the Biobank Act has raised speculation of the scope of data one might obtain from a biobank and discussion of the possible risks and harmful implications it may have on data subjects. In the age of biomedical big data, right to access may in practice cover a very large amount of information that researchers are obligated to return into the biobank, for example raw genomic data. This in itself raises various ethical and legal concerns of responsibility as to how such data is interpreted and used once disclosed to the data subject.

The second, and possibly more troublesome part of access to information, relates to findings with clinical relevance to the data subject. According to Section 39.2 of the Biobank Act, "a registered person has the right, upon request, to receive information regarding his or her health as determined based on a sample." In such a case the registered person shall be offered a possibility to know the meaning of such information. The right to information regarding a person's health derives from the Biomedicine Convention, according to which "everyone is entitled to know any information collected about his or her health (Article 10.2). During the national legislative process, the right to be informed was limited to incidental findings, which are clinically significant (actionable), thus excluding random research results, e.g. raw genomic data which has not been properly analysed, arranged into knowledge or interpreted for health purpose of the data subject. Hence, the legislator has chosen both to empower the data subject by granting a choice for seeking information, and on another hand has taken a very paternalistic stance in protecting the data subject from potentially harmful information if he or she has chosen not to consent (right not to know).

The duty to disclose clinically relevant findings has turned out to be a challenge for biobanks to implement, although to the writers' knowledge no such requests have so far been made. However, concerns are justified as a majority of Finnish data subjects have consented to receiving findings, generated from a biobank. In absence of a nationally coherent policy regarding returning of clinically significant findings, each Finnish biobank is faced with the challenge of independently defining the relevant findings and the methods for disclosing such information to the registered person. The biobanks have elaborated their individual feedback policies and addressed a range of issues such as re-contact, carrying out new tests and returning individual data. There is still significant uncertainty as to what type of actionable data should be returned and no guidance has been provided for this assessment. Transferring responsibility to researchers via Material Transfer Agreements (MTA) has turned out to be a viable option as researchers could be required to submit an evaluation to the biobank of potential findings and a description of the methods of observing and processing them in the research project.

Then the biobank could, with the help of a multidisciplinary scientific advisory group, deliberate the need to return such findings to the biobank. The obligation to define the data to be disclosed is hardly a responsibility biobanks voluntarily want for themselves and have expressed wishes for a policy to be drafted in the context of the national genome strategy (Report of the biobank legislation steering group 2015). So far, none of the biobanks have separately addressed the issue of returning data to paediatric populations or families of the data subjects.

As described above, the right to receive information regarding one's health has been experienced as problematic in practice mainly because it involves a corresponding duty for biobanks. Sample analysis generates large volumes of individual genetic data (incl. individual research results and incidental findings), but the analytical validity, clinical validity and clinical utility of the data isn't always clear (Thorogood et al. 2014). Limited understanding and difficulties in interpreting the data may become overwhelming for patients, practitioners and the health care system. It also brings about concerns over liability if a biobank doesn't detect the clinically relevant data. Not to mention the liability issues resulting from risks generated by breaking the double coding system for assessment and communication of findings. It is also noteworthy that the duty to disclose clinically relevant information legally obligates biobanks, rather than places an ethical duty to act for the researchers using the biobank's samples, even though research results are commonly produced in research projects outside of the biobank. Additionally, research laboratories aren't in all cases accredited for clinical purposes and don't use methods as accurate as those in a clinical laboratory, which is a prerequisite for giving clinically valid results to data subjects. The Additional Protocol for biomedical research (Article 27) states the duty of care and requires that information should be offered within a framework of health care or counselling. This is problematic because biobanks are a research infrastructure, which don't directly have anything to do with a data subject's clinical care and don't have a possibility to verify if the data subject has a current relationship with a clinic. Possibilities to offer genetic counselling are limited due to scarce resources.

The Finnish government has just recently published a national genome strategy (Ministry of Social Affairs and Health 2015a). Biobanks are increasingly seen as mediators of personalized health as they tend to generate vast amounts of health related data. As a part of execution of the strategy, a specialist panel or group will presumably be formed to draft a policy for returning clinically relevant information to the public. This move would in effect possibly extend the legal right to receive health-related information from biobanking regulation to a wider principle of genomic data policy. The working group behind the recent strategy has proposed establishment of a national Genome Centre, which could act as a governing body to develop criteria for contacting individuals. This move would not remove the need to address the basic questions of what information ought to be deemed returnable to individuals and where should thresholds of risk be drawn with regard to findings in genomic research. A basic question behind these issues is what role the state should assume in the management of data derived from biobank research. Issues related to incidental findings and return of individual results are becoming increasingly a state wide concern in that for many countries, such as the Nordic countries the

data available for analysis is becoming increasingly state wide to include major parts of the population. As a result, one needs to ask whether it is ethically suitable for a welfare state to disperse choices to biobanks and individuals or whether state authorities should play a stronger coordinating role in the governance of data derived from such large data sets.

In the following section we will discuss the ethical dilemma that EU member states are faced with in relation to the governance of big data in biobanking. We will particularly focus on the challenges that biobanking activities and their administration face with relation to the way healthcare services are organized within the welfare state system.

3 Governing Big Data Findings in Biobanking

A central feature of recent biobanking governance discussions has focused on the ways in which the ever increasing amount and significance of big data related to health will be governed, especially if it contains information that is of actionable relevance to individuals. Hancher and Moran (1989) have suggested that the notion of regulatory space is useful as a conceptual tool to explore the actors within different regulatory regimes (see also Kaye et al. 2012). The notion of regulatory space suggests that state power is not concentrated in one authority or actor, namely the state in this case. Rather it is dispersed among various state actors and bodies, which may include non-governmental actors as well. Therefore, in order to understand the regulatory process, one has to look at all the partners and actors that are involved throughout the regulatory process. Within the Finnish context this space is for the most part populated by state authority, state bodies and experts, ranging from regulatory officials, and national ethics committees to medical and genetic researchers who have a stake in biobanking. This also invariably includes the biobanks and their managers as well. The broader context of this regulatory space, however, can also be seen to be located within the Nordic welfare state system. Within this system there is a general ethos of universal access to healthcare services and equality in the sense that the system should strive to provide the same services to all citizens regardless of age, sex or place of residence. Although there have been substantial reductions in the ways in which healthcare services have been provided, as well as an increase in the offerings from the private sector, the same ethos of universality and equality still serves as a guiding principle. Still, different biobanks operate in different regulatory spaces, although most biobanks are still guided by a general principle of supporting research into the causes of disease.[6] A number of cases from around the world highlight the challenges that biobanks are facing in managing big data as it relates to individuals and privacy concerns.

[6]These differences are explained by the institutional roles that they have; some biobanks are based on specific research into particular cancers, such as prostate cancer, while other biobanks are hospital-based diagnostic collections based on millions of samples. The regulatory space in which research groups and hospitals operate in are vastly different.

Recently in Iceland, the CEO of deCode Genetics, Kari Stefansson, noted that they had data on all the breast cancer gene BRCA2 carriers in Iceland. This had been possible by having samples from some Icelanders and being able to impute the rest based on the detailed genealogies that the country had maintained. DeCode, however, only have this information in a coded format so it was not able to contact and warn the individuals themselves since Iceland's biobanking legislation prevents deCode from having direct access to patient identities. As a consequence, the Icelandic government is reviewing its policies regarding return of individual research results and findings to individuals who might be affected by such information.

Similarly, UK Biobank, which has had a strict policy of not returning results of its studies to participants have had to re-evaluate and change this policy in relation to a study which also uses MRI scans since the radiologists performing the scans felt that not notifying participants of an incidental finding, such as a tumour, would have been medically unethical. According to UK Biobank:

> UK Biobank is working with social scientists and health economists to gain a better understanding of the risks and benefits associated with providing feedback of potentially serious incidental findings to UK Biobank participants during the imaging pilot study. In some cases, these incidental findings can have serious health implications; in others, the medical implications are less clear, and many potentially serious findings may – after further investigation or the passage of time – turn out not to be of concern after all. The impact that feedback of information about potentially serious incidental findings has on participants has not been well researched. This work is important because there is currently no consensus in the research community on which (if any) incidental findings should be fed back and the best methods for doing this. (UK Biobank 2015)

What both these cases indicate is that biobanking is raising many ethical questions surrounding the situations under which biobanks would be obligated to report findings to research subjects, as well as the methods through which this is best done. In both cases issues of managing various forms of data derived from biobanking and subsequent studies are managed on a case-by-case basis without coordination among other biobanks. This leads to situations in which some biobanks may adopt on policy regarding governance of data, while others may adopt another policy. Given that biobanking infrastructures are being increasingly constructed with the idea of linking ever larger sets of data from samples and other data registers together, it would seem surprising then that citizens and patients in one political context would be placed in a different status as opposed to others with regard to the types of information that they would be able to benefit from based on that research. A first step in this process is, to develop some form on national coordination as to how big data ought to be managed.

A number of recent studies have argued that there is a moral responsibility among researchers to inform individuals of the likelihood of a serious condition (Miller et al. 2008; Fernandez et al. 2003). Finland's biobanking law sought to find a balance with regard to granting researchers broad rights in relation to collection and use of coded samples by also giving research subjects the right to inspect what their samples were being used for, as well as the right to request information

regarding the significance of said research for their health. This approach follows what some commentators have called "dispersed ethics" where a common feature of high-tech biomedicine is that ethical choices are dispersed to individuals (Hélen 2010, 30). According to this argument, the state retreats to a position of technical decision making whereas the individual and her relationship to new technologies, such as whole genome sequencing, is appropriated as a matter of individual choice. Technological development and ones participation in high-tech medicine through samples donation is seen as an important nexus in which the participant is empowered and expected to seek out information regarding their own health and then act upon that information accordingly. Burke et al. (2014, 107) have, however, argued that the weight of bioethical and researcher opinion argues against granting research subjects an unrestricted right to demand return of individual research results. This point is also picked up by some of the people that we have interviewed. Some hospital administrators, for example, expressed a concern that under the current legislation only younger and more informed patients would actively seek out information on what had been studied using their samples. This, according to them, would lead to health inequalities among different sectors of the population. Furthermore, a number of interviewees also noted it was also unclear who should be the person to impart the information and on what basis should decisions of what is relevant information be made.

Such concerns among interviewees raises a number of important ethical questions related to the governance of big data. *First*, in relation to the notion of equality and universality, the hospital administrator ponders whether certain segments of the population will be more active than others in seeking out genomic information and its relevance to their health. This has the potential to create more inequality since many segments of the population might not be as active in seeking out information as others. In fact, most people might not even be aware that they have the right to do so. In contrast to the notion of *therapeutic misconception* (Zawati and Knoppers 2012; Stjernschantz et al. 2009; Appelbaum et al. 1982), one could begin to talk about the notion of *therapeutic non-conception*, whereby people are not aware of how their samples and information are being used. A number of studies have shown that most people do not possess sufficient information of what a biobank is (Tupasela et al. 2010), so it is highly unlikely that anyone would be actively seeking information regarding their health based on a sample they may have in a biobank. Furthermore, the people who are at greatest risk of becoming chronically ill are usually those who are socially disadvantaged, and who tend not to seek help from the healthcare services in the first place.

Second, the excerpt highlights the challenge that biobanks may have in trying to interpret what is of significance to a research participant from the vast amount of information that may be contained in the research. Biobanks have been developed as an infrastructure and processing service for researchers, not as experts in the research results themselves. Even more significant is the fact that most biobanks may not have a qualified genetic counsellor on staff to interpret the results, nor to discuss them with the research participant. This also raises the issue of biobank context in relation to return of information.

Third, institutional heterogeneity is also a major challenge in this approach. As mentioned above, there are a number of different types of biobanks which have been set up in Finland. Some are based on large cohorts in research organisations, some are clinical collections in hospitals, while others may be based on disease specific collections and research programs into prostate cancer, for example. As a result, these different instantiations of biobanks may have very different contact zones with their research subjects or patient populations. Clinical collections collected for the treatment and study of specific cancers will have a vastly different interface with the population they are treating and studying than large population cohorts, and thus their ability to return results and trust their clinical validity and significance may be very different. The same goes for the patients and research subjects, patients being treated for cancer may be much more active in seeking information than research participants who have donated a sample for a cohort study.

The regulatory space that has been created with the current form of biobank legislation in Finland places a high degree of emphasis on individual responsibility in seeking out information, which most people may not even be aware of. As a result, although it respects the notion of personal autonomy and rights it is highly inefficient from a public health perspective and may lead to increased inequalities within the population in terms of access to important information. In the following we will discuss an alternative to the current legislative format and highlight some of the benefits that it may have in relation to the development of a national genome strategy. This follows a strong demand from the Finnish biobanking community, who see a need to change the current legislation (Lautala 2013).

4 State Responsibility and Accountability in Biobank Research

During the past decade, Finland – like many other European countries – has made strives to increase the role of the private sector in the delivery of healthcare services. Within this process, individual consumer choice has been given increased visibility and emphasis as a driver of markets and policy-making (Häyrinen-Alestalo et al. 2009). Biobanks have been seen as an important nexus around which new healthcare markets will develop and produce new treatments and services for consumers (OECD 2001, 2005; EU Workshop 2003; Sitra 2014). The existing biobank act in Finland follows in many ways the logic of market operations when it comes to responsibility and accountability in the delivery of healthcare services. It has drawn on a form of dispersed ethics, whereby the responsibility for seeking health related information is located within the individual consumer. Furthermore, it disperses ethical decision making to the biobanks themselves in that the law expects each biobank to be able to decide which information is relevant and significant for the individual. The lack of centralized coordination in this matter is a significant departure from the basic principles which are ascribed within the traditional Finnish healthcare system and which have met with problems with the current form of the biobank legislation.

As a result of the tension between the welfare state ethos and practice there has emerged a need for policy makers and regulators to revise the act (see Ministry of Social Affairs and Health 2015b). This regulatory revision process operates within a regulatory space in which the limits of responsibility and accountability are being defined. One of the main challenges faced is how to assure equality in a system which has dispersed responsibility to both biobanks and individuals.

The significance of the role and position into which the state is placed indicates a number of salient features about biobanking and healthcare. *First*, it is highly unlikely that the state will be able to disperse ethical decision making to individuals through the use of so-called consumer choice in the healthcare market. *Second*, the state is likely to have to operate as a coordinator of biobanks in developing a national policy or guidelines for the management and governance of big data in biobanking, particularly as it relates to incidental findings and individual research results. The challenges which will emerge will relate to the ways in which variants of uncertain significance will be managed at a national level, as well as which criteria and conditions will be applied to determining their national health significance and how the information will be disclosed. *Third*, the state, along with the biobanks will need to develop a strategy through which big data will be managed in the future. Given that research using tissue samples are conducting whole genome or exome sequencing, there will emerge a need to find a common standard and policy as to the ways in which information on patients will be stored and managed. This will be necessary, not only as a cost minimizing process, but it will also be necessary to keep track of what information is produced and stored concerning each patient. A patchwork system where each biobank will develop its own policies will lead to healthcare inequalities between patients and research subjects in different regions, and this would be counter to the rules and goals of the state as a guarantor of equal access and position within the state-led healthcare system.

We began this chapter by highlighting the role of the EU Data Protection Directive in defining the different rights and obligations while processing personal data. In the Finnish biobanking field current debates focus on the rights and responsibilities relating to access to data, even raw unanalysed genomic data. If it includes information which may have relevance for the data subject's health, he or she has a right to know and a biobank has a corresponding legal duty to disclose. However, this legal duty has been subordinated to the activity of the data subject. Alongside the legal duty, there remains an ethical duty to disclose, but even this duty is subject to the consent of the data subject. As a result, both duties seem to be dependent of the responsibilities the legislator has addressed to an individual person. The choices taken by the Finnish legislators reflect the current trend of the big data era: empowering people. However, by reflecting on the examples given in this chapter, one might ask: how much power does empowerment have, if the value of the information one has a right to receive is close to zero, minimal or uncertain? Should governments opt to give out all information in the name of empowerment? Or should they take a careful stance and give out limited information? If so, should the data be classified as high or low risk, as a form of data? Or should the data be valued by the relevance to an individual's health?

Europe continues to anticipate the wordings of the future data protection regulation, classification for data and the implications for both biobanking and larger data sharing infrastructures. Effective responses to big data frameworks and returning individual data can be made only if we can agree on common interpretations of international rules and regulations. In doing so it is important to recognize the different interests and values of the stakeholders operating in the big data sphere, which includes strong economic forces as a counterpart to individual rights. The existing local and European norms have been shown to be inconsistent and need to evolve taking into account the different shades and nuances of global translational medicine and abandon old dichotomies between research and clinical care. The Biomedicine Convention provides a legal framework and four leading principles upon which to build and promote policy solutions for integrating data deriving from biobanks and other data sharing infrastructures into the healthcare setting. In light of the principles, it is necessary for governments to consider for example: (1) who needs protection and from what? What are the risks to dignity, identity and integrity? (2) What does equal access to healthcare mean as a legal and social right? (3) How do we define non-discrimination both in the public and private sector? (4) How can the challenges we face be met without excessively affecting the core principles of the Nordic health care system?

Within the Nordic research systems there has also been a long tradition of ensuring a broad range of rights for the scientific research communities to do high quality research using a broad range of population registers, health care data, as well as – increasingly – tissue samples. These research infrastructures increasingly rely on the ability to access and use big data that is collected by the state or various healthcare institutions and organisations, but also rely on public support for their long-term sustainability (Tupasela et al. 2015). Placing limits on the ability to collect sensitive data on populations, as well as limiting the ability of researchers to access it would have serious implications in the development of new diagnostics and treatments. In relation to the management of big data relating to biobanking it is important that there develops coordinated systems at national levels which can take responsibility for overseeing the activities of a broad range of heterogeneous actors to ensure that the work that they do is not wasted, replicated or the benefits of which may end up being unevenly distributed across the population. Given that the notion of equal access and fairness can be considered important principles in the healthcare sector, it would seem appropriate that such principles also be applied to the organization of biobanking activities as well. It is important that biobanks are not left to their own devices in relation to developing governance structures and policies for the management of big data. Given that biobanks, especially in the Nordic countries (cf. Hoeyer, chapter “Denmark at a Crossroad? Intensified Data Sourcing in a Research Radical Country”, in this volume), are increasingly able to integrate data from a broad range of resources, such as national registers, it is imperative that a coordinated effort and governance structure is developed to ensure that the principles of equality and fairness are not disregarded. It is important that

there is a common effort to develop policies to decide what information is deemed relevant, as well as irrelevant in terms of developing national genome policies for better health of the whole population.

References

Appelbaum, P.S., L.H. Roth, and C. Lidz. 1982. The therapeutic misconception: Informed consent in psychiatric research. *International Journal of Law and Psychiatry* 5: 319–329.

Aromaa, Arpo, Veikko Launis, and Salla Lötjönen. 2002. *DNA-näytteet epidemiologisessa tutkimuksessa. DNA ja Epidemiologia-työryhmä*. Helsinki: TUKIJA/ETENE. [DNA samples in epidemiological research].

Bledsoe, Marianna, Ellen Wright Clayton, Amy McGuire, William Grizzle, Pearl O'Rourke, and Nikolajs Zeps. 2013. Return of research results from genomic biobanks: Cost matters. *Genetics in Medicine* 15: 103–105.

Burke, Wylie, Hilary Burton, Alison Hall, Mohamed Karmali, Muin Khoury, Bartha Knoppers, Eric Meslin, Fiona Stanley, Caroline Wright, and Ronald Zimmern. 2010. Extending the reach of public health genomics: What should be the agenda for public health in an era of genome-based and "personalized" medicine? *Genetics in Medicine* 12(12): 785–791.

Burke, Wylie, Barbara Evans, and Gail Jarvik. 2014. Return of results: Ethical and legal distinctions between research and clinical care. *American Journal of Medical Genetics* 166C: 105–111.

Cambon-Thomsen, Ann, Emmanuelle Rial-Sebbag, and Bartha Maria Knoppers. 2007. Trends in ethical and legal frameworks for the use of human biobanks. *The European Respiratory Journal* 30(2): 373–382.

Council of Europe. 1981. Convention for the protection of individuals with regard to automatic processing of personal data. ETS 108. Strasbourg, 28.I.1981.

Council of Europe. 1997. Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine. ETS 164. Oviedo, 4.IV.1997.

Council of Europe. 2005. Additional protocol to the convention on human rights and biomedicine, concerning biomedical research. ETS 195. Strasbourg, 25.I.2005.

Directive 95/46/EC of the European Parliament and 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 (Data Protection Directive).

Dute, Jos. 2005. The leading principles of the convention on human rights and biomedicine. In *Health law, human rights and the biomedicine convention. Essays in honour of Henriette Roscam Abbing*, ed. J.K.M. Gevers, Ewoud H. Hondius, and J.H. Hubben, 3–12. Leiden/Boston: Martinus Nijhoff Publishers; Leiden: Brill.

Eric, Juengst, Flatt Michael A., and Richard A. Settersten Jr. 2012. Personalized genomic medicine and the rhetoric of empowerment. *Hastings Cent Rep* 42(5): 34–40. doi:10.1002/hast.65.

European Commission. 2013. *Commission Staff Working Document – Use of '-omics' technologies in the development of personalised medicine*. Brussels: European Commission.

EU Workshop. 2003. Biobanks for health – Optimising the use of European biobanks and health registries for research relevant to public health and combating disease. Report and recommendations from an EU workshop held at Voksenåsen Hotel, Oslo, 28–31 Jan.

Fernandez, Conrad, Eric Kodish, and Charles Weijer. 2003. Informing study participants of research results: An ethical imperative. *IRB: A Review of Human Subjects Research* 25(3): 12–19.

Frank, Lone. 2000. When an entire country is a cohort. *Science* 287(5462): 2398–2399.
Genomics England. 2015. Genomics England and the 100 000 Genomes project. http://www.genomicsengland.co.uk/the-100000-genomes-project/
Hallinan, Dara, and Michael Friedewald. 2015. Open consent, biobanking and data protection law: Can open consent be 'informed' under the forthcoming data protection regulation? *Life Sciences, Society and Policy* 11: 1.
Hancher, Leigh, and Michael Moran. 1989. Organising regulatory space. In *Capitalism, culture and economic regulation*, ed. L. Hancher and M. Moran, 271–299. Oxford: Oxford University Press.
Häyrinen-Alestalo, Marja, Ville Mälkönen, and Pekka Valkama. 2009. *Markkinamekanismit julkisissa palveluissa*. Tekesin katsaus 253. Helsinki: Tekes [Market mechanisms in public services].
Hélen, Ilpo. 2010. Technics over life: Risk, ethics and the existential condition in high-tech antenatal care. *Economy and Society* 33(1): 28–51.
Hoeyer, Klaus. 2012. Size matters: The ethical, legal, and social issues surrounding large-scale genetic biobank initiatives. *Norsk Epidemiologi – Norwegian Journal of Epidemiology* 21(2): 211–220.
Kaye, Jane, Liam Curren, Nick Anderson, Kelly Edwards, Stephanie M. Fullerton, Nadja Kanellopoulou, David Lund, et al. 2012. From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews Genetics* 13(5): 371–376.
Knoppers, Bartha Maria, Yann Joly, Jacques Simard, and Francine Durocher. 2006. The emergence of an ethical duty to disclose genetic research results: International perspectives. *European Journal of Human Genetics* 14: 1170–1178.
Lautala, Tiina. 2013. Biopankkilaki ei taivu käytäntöön ongelmitta. *Suomen Lääkärilehti* 68(50–52): 3300–3302.
Miller, Fiona, Mita Giacomini, Catherine Ahern, Jason Robert, and S. Sonya de Laat. 2008. When research seems like clinical care: A qualitative study of the communication of individual cancer genetic research results. *BMC Medical Ethics* 9(4): 1–12.
Ministry of Social Affairs and Health. 2015a. *Parempaa terveyttä genomitiedon avulla – Kansallinen genomistrategia työryhmän ehdotus*. Tampere: Juvenes print. [National Genome Strategy].
Ministry of Social Affairs and Health. 2015b. *Biopankkilainsäädännön ohjausryhmän väliraportti 2015*. Helsinki: Sosiaali- ja terveysministeriö. [Report of the biobank legislation steering group].
Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.
National Supervisory Authority for Welfare and Health. 2015. *Valviran ja tietosuojavaltuutetun toimiston yhteinen ohjaus koskien näytteeseen ja siihen liittyvän tiedon siirtämistä biopankkiin, tutkimusnäytteiden jatkokäytön turvaamista ja tunnisteellisten näytteiden sekä niihin liittyvien tietojen luovuttamista biopankista*. 24.3.2015. Dnro 2281/06.01.05.00/2015. [Valvira guidance letter].
OECD. 2001. *Biological resource centres: Underpinning the future of life sciences and biotechnology*. Paris: OECD.
OECD. 2005. *The bioeconomy to 2030: Designing a policy agenda*. OECD International Futures Programme. Available at http://www.oecd.org/dataoecd/47/19/35532457.pdf
OECD. 2013. Integrating personalised medicine into health care: Opportunities and challenges. In *ICTs and the health sector: Towards smarter health and wellness models*. OECD Publishing. http://dx.doi.org/10.1787/9789264202863-8-en
PACITA. 2014. *Future panel on public health genomics expert working group reports* (6 Feb 2014). Parliaments and Civil Society in Technological Assessment. Available at http://www.pacitaproject.eu/wp-content/uploads/2014/02/WP-5-EWG_reports_on_Public_Health_Genomics_-_DEF_6_Feb_2014.pdf.
Personal Data Act (523/1999).

PwC. 2005. *Personalized medicine: The emerging pharmacogenomics revolution*. Global Technology Center: Health Research Institute. Available at http://farmagenomica.altervista.org/DOCUMENTI/PWC.pdf.

Sitra. 2014. Tax income, investments, jobs – From biobanks? *Sitra Blog* 30.7.2014. Available at http://www.sitra.fi/en/blog/tax-income-investments-jobs-biobanks. Downloaded 30 June 2015.

Soini, Sirpa. 2013. Finland on a road towards a modern legal biobanking infrastructure. *European Journal of Health Law* 20(3): 289–294.

Stjernschantz, Forsberg, Mats Hansson Johanna, and Steffan Eriksson. 2009. Changing perspectives in biobank research: From individual rights to concerns about public health regarding the return of results. *European Journal of Human Genetics* 17: 1544–1549.

Thorogood, Adrian, Yann Joly, Bartha Maria Knoppers, Tommy Nilsson, Peter Metrakos, Anthoula Lazaris, and Ayat Salman. 2014. An implementation framework for the feedback of individual research results and incidental findings in research. *BMC Medical Ethics* 15: 88.

Tupasela, Aaro. 2015. Tensions between policy and practice in Finnish biobank legislation. *Biopreservation and Biobanking* 13(5): 379–381.

Tupasela, Aaro, and Karoliina Snell. 2012. National interests and international collaboration: Tensions and ambiguity among Finns towards usages of tissue samples. *New Genetics and Society* 31(4): 424–441.

Tupasela, Aaro, Sinikka Sihvo, Karoliina Snell, Piia Jallinoja, Arja Aro, and Elina Hemminki. 2010. Attitudes towards the biomedical use of tissue sample collections, consent and biobanks among Finns. *Scandinavian Journal of Public Health* 38: 46–52.

Tupasela, Aaro, Karoliina Snell, and Jose A. Cañada. 2015. Constructing populations in biobanking. *Life Sciences, Society and Policy* 11(5): 1–18.

UK Biobank. 2015. *Incidental findings research*. http://www.ukbiobank.ac.uk/incidental-findings-research/.

Wolfe, Susan, et al. 2012. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. *Genetics in Medicine* 14(4): 361–384.

Zawati, Ma'n H., and Bartha Maria Knoppers. 2012. International normative perspectives on the return of individual research results and individual findings in genomic biobanks. *Genetics in Medicine* 14(4): 484–489.

Zika, Eleni, et al. 2010. *Biobanks in Europe: Prospects for harmonisation and networking*. Luxembourg: JRC Scientific and Technical Reports.

Big Data, Small Talk: Lessons from the Ethical Practices of Interpersonal Communication for the Management of Biomedical Big Data

Paula Boddington

Abstract Biomedical big data raises various ethical issues, many of which concern tensions between the public sharing of information and the private control of personal information, and the status of the individual data subject. Close attention to the intersection between issues in epistemology and in ethics is needed, and this chapter outlines divergent models of the transmission of information which give rise to different understandings of the ethics of communication. This chapter also draws on work in social sciences which examines a parallel area where ethically similar issues arise, the communication of personal and familial medical and genetic information. Analysis of a body of research draws attention to the situated and personal communication of knowledge, explaining how this generates ethical considerations which may clash with impersonal or system-driven understandings of data, and showing how individuals might display alternative ways of understanding their ethical responsibilities. Important ethical questions arise such as how, whom, and when to tell. Channels of communication may guide questions of ethical responsibility. These considerations emphasise the importance of context and are used to extend Nissenbaum's notions of contextual integrity. The chapter also examines the question of the disempowerment of the data subject, and suggests that the changing patterns in the dissemination of biomedical data may provide individuals and groups with ways of acting which may help to redress fears of the disempowered data subject.

1 Introduction

Biomedical big data is often said to be destabilising in nature, and in the area of ethics this potential to destabilise should be of particular concern. It should alert us to the need to cast our net widely in considering exactly how we understand the ethical issues involved, and what we might or might not do about them; in an area which has been destabilised, who knows what we might find. We should think

P. Boddington (✉)
Department of Computer Science, Oxford University, Wolfson Building, Oxford OX1 3QD, UK
e-mail: paula.boddington@cs.ox.ac.uk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_13

widely both in terms of looking further afield at work on ethical questions in other areas which might prove illuminating for biomedical big data. We should also take sufficiently seriously the notion of destabilisation to ask if we are framing the ethical issues in the best way. What fundamental assumptions might we be making, about big data, and about ethics, which might need to be questioned? Where might our assumptions need to be adjusted, supplemented, or replaced?

This chapter suggests that lessons can usefully be learned from a variety of sources, and in particular attempts to learn lessons for big data from examining normative questions in personal communication. Genomics research, where similar questions of the handling of large amounts of data arise, is examined, as well as work on ethical issues concerning the handling of information in clinical genetics. We will also briefly consider how lessons currently being learned from the use of electronic data gathering and analysis and mobile technology in health care may be applied more broadly to thinking about ethical issues in biomedical big data. The chapter draws upon some empirical findings and examines how these may lead us to reconsider underlying conceptual frameworks, especially in the intersection between ethics and epistemology. Much of the discussion will refer to work in social sciences. The questions to be considered are highly complex, and it should be noted at the outset that the aims of the current chapter must be correspondingly modest, to suggest some possible ways forward for future work.

1.1 Assumptions About Ethics

It will be useful at the outset to challenge some often unspoken yet implied assumptions about how to tackle applied ethical issues, since this will have particular relevance to how we proceed and what conclusions we might draw. Firstly, the place of institutional and regulatory frameworks for ethics must be considered. There is obviously a great need in various areas of professional and public life to put in place robust, well thought out, useable and meaningful regulations. But to consider that regulations on their own will 'solve' the issues is of course misguided. Regulations need the backing of an institutional and societal culture which has the will and ability to implement them, which can recognise when rules should be applied, which is open, and allows challenge and scrutiny. Institutions and the individuals working in them need to have sufficient integrity and sufficient freedom to make institutional regulations work well. Without this, regulations may be useless. Indeed, they may be worse than useless, in giving the appearance that all is well. My favourite example is the magnificent sounding code of ethics extolling values of respect, integrity, communication and excellence in place at the failed company Enron, which collapsed spectacularly in 2001 taking with it the livelihoods of hundreds of thousands of ordinary people. Trouble is, few if any took one scrap of notice of a code of ethics which exhorted people to integrity in a company whose internal culture appeared to be dominated by individualism and by short sighted

greed (McLean and Elkind 2004). So, even if a comprehensive and robust code of ethics for the use of biomedical big data can be drawn up and implemented, it will not alone be enough.

In happy contrast, there are more heartening examples of researchers who have demonstrated their integrity in thinking proactively when their research showed that the data they were using was unexpectedly powerful, for example when aggregated with other data sets. Examples include caution exercised in combining data sets which threatened anonymity by successfully predicting surnames (Gitschier 2009); pre-emptive advice taken when work demonstrated the potential to determine the presence of individuals in large groups of aggregated genomic data (Homer et al. 2008); and warnings issued and steps taken when work revealed that James Watson's ApoE gene status could be accurately conjectured from publicly available data (Nyholt et al. 2009). ApoE is a gene that codes for a protein implicated in the development of Alzheimer's disease: certain versions of the gene give a higher risk of developing late onset Alzheimer's disease (For a discussion see Boddington 2012, p. 33.). This demonstrates the need for committed, resourceful individuals with the foresight and motivation to anticipate issues. Given the disruptive nature of big data, this need is not likely to dissipate any time soon. The need for integrity and virtuous conduct in implementing ethical regulations has also been particularly apparent in research in the social sciences, where the fine-grained nature of the work, the propensity to make unexpected discoveries in the course of one's research, and the frequent exploration of novel research methods, has meant that researchers may need to be particularly alert to ethical problems which might outstrip the capacity of regulatory frameworks (Banks 2015). These features are not confined to the social sciences, of course, and are reproduced in research concerning biomedical big data.

Secondly, what kind of answers can we expect in ethics? Aristotle warned when writing of ethics that 'Our discussion will be adequate if we make things perspicacious enough to accord with the subject matter, for we would not seek the same degree of exactness in all sorts of arguments alike, any more than in all the products of different crafts' (Aristotle 1999, p. 2). However, perhaps driven by the very need of regulation to produce answers, there is often an assumption that ethics ought to be like a problem in maths or in logic: we must have answers to everything, and we must have them now. Of course this is driven by the importance of ethical questions and the desirability of having answers to ethical quandaries. But if this is based on a simplistic idea that it will be possible in all cases to provide a unique solution, we must exercise caution. On some accounts of ethics, there is no such thing as an unresolvable ethical dilemma. Of course in some cases, we may lack the knowledge to be able to work out what to do, but in theory, some claim, there will be the best thing to do, all things considered. On other accounts of ethics, there are genuine dilemmas where, whatever you do, you end up doing something wrong (Nagel 1979; Williams 1981). Given the uncertainties and the destabilising nature of biomedical big data, it may be a prudent strategy to act as if the former view were true, in the hope of finding a solution, but to bear in mind that the latter view might be more realistic. At the very least, it's vital to think holistically, as focus too narrowly on addressing one ethical issue can reduce the

capacity to respond effectively to other ethical problems: for example, requiring certain levels of reciprocity between data sharers may help to address some ethical questions, but might inadvertently exclude researchers from resource-poor countries (Boddington 2012, p. 195 ff). If there are ethical dilemmas which remain, this again gives reason to think carefully and creatively about how unsatisfactory situations may be ameliorated or how we may learn to live with them. Again, this will involve thinking more widely than simply focusing on regulations.

There may certainly be a 'catch up' phase as we grapple with how to adjust to the world of biomedical big data and how best to live with it. But as big data rushes ahead, the very urgency of the ethical questions that arise should warn us to tackle them well, rather than concentrating on tackling them quickly.

2 Ethical Questions About Biomedical Big Data

In a recent paper, Mittelstadt and Floridi (2016) review the literature on the ethics and morality of biomedical big data and summarise eleven areas of concern. Five were found to be frequently mentioned in the literature: informed consent, privacy including anonymisation and data protection; ownership; epistemology and objectivity; and big data divides including resource inequality. Six were flagged as lacking attention but nonetheless meriting concern and future work: group level harms; the importance of epistemology in assessing the ethics of big data; the changing nature of fiduciary relationships in the light of big data; differences between academic and commercial contexts in the use of big data; intellectual property and the analysis of aggregated data sets; and meaningful access rights for resource-poor individuals.

It would be over-ambitious to attempt to address too many of these questions in this chapter, but it will be noted in passing that some of the issues flagged as needing further consideration have had some attention from those working on ethical issues in genomics. For example, right from its conception, the HapMap consortium, which collected material from individuals from certain population groups for referencing purposes in genomics research, addressed ethical issues including group level harms and benefit sharing (HapMap 2004; Rotimi et al. 2007; Austin 2002). Other work has addressed how changing fiduciary relationships and responsibilities might arise from the disruptive effects of aggregating genomic data (e.g., Cassa et al. 2008); how to protect particular populations including disadvantaged groups (Knoppers 2000; Beskow et al. 2001); and benefit sharing of research both with participants and with researchers from resource-poor countries (Mascalzoni et al. 2008). Other resources ripe for tapping for considering ethical issues in biomedical big data include work on social science research which explores issues of consent, the individual and the group. For example, ethical difficulties arise in obtaining informed consent, which is premised upon treating *individuals* qua individuals with their own autonomy, when the research question is premised upon interest in that individual as a member of a particular *group* (see e.g., Atkinson 2009).

Mittelstadt and Floridi quite understandably lament that despite their search covering 68 papers on ethics of big data, the literature leaves gaps in empirical research and in deep conceptual analysis, partly because many were editorials or short opinion pieces (2016). This is doubtless partly due to the newness of the topic, but also perhaps illustrates a 'big data' style problem – the pressure to publish in academia creates a speed of output which may militate against depth, and the very form of academic journals also pushes ethical debate into certain categories and word limits which, just as constraints of data storage and analysis may shape what is produced, and, together with the newness of the topic, again may limit the outputs. It also doubtless reflects the difficulty of constructing ethical responses to big data, where individual actors are increasingly rendered seemingly impotent as they are seemingly construed as mere data points and profiles. The dominant model in medical ethics at least in Anglo-American contexts focuses firmly on the individual (Beauchamp and Childress 2009). If big data raises questions about the nature of the individual, and seems to threaten to reduce individual power, no wonder there are difficulties in addressing ethical questions, since a notion at the hearts of contemporary medical ethics is itself in danger of disruption.

Many of the ethical questions which Mittelstadt and Floridi discuss concern, in broad terms, the relationship of individuals to groups: of the individual data subject to the data custodian, the data analyser, and the data user, often represented by powerful groups; and additionally, questions about the relationship of less powerful groups to more powerful groups. Indeed, it could be said that 'Considered together, the emerging picture is of data subjects in a disempowered state, faced with seemingly insurmountable barriers to understanding who holds what data about them, being used for which purposes' (Mittelstadt and Floridi 2016: 331). It is this disempowered state in particular that I wish to examine and, to an extent, begin to challenge. Biomedical big data poses obvious problems in how we see the subject of that data. Note that all of these questions are mediated through questions about knowledge, and these questions in the relationship between ethics and epistemology are vital. It is to these broad issues that this chapter looks. We also need to consider if the very ways in which we ask ethical questions about these problems understand the individual subject adequately.

3 Preliminary Thoughts: Epistemology and Ethics in Biomedical Big Data

Many of the ethical questions of biomedical big data are mediated through the issue of knowledge, or more broadly, the issue of the position of individuals (and groups) in relation to data. Indeed, once the problems are presented in this broad way, the relevance of epistemology to ethical issues in biomedical big data become quite obvious, since how individuals stand in relation to information is, in broad terms, tautologically an epistemological issue. It is here then that there is a pressing need

for deeper conceptual analysis and indeed further empirical work. This chapter will argue for the need to understand epistemological issues not simply in terms of what is known or not known, and how it is known or not known, but also in much richer terms.

This section thus reviews the issues in ethics and epistemological issues that need attention, and also discusses some common but problematic assumptions about epistemology and ethics in biomedical big data. These discussions will form an essential backdrop to later discussions.

Firstly, there is a tendency to see big data as 'objective' and hence, as beyond question (Crawford 2013). The alleged objectivity of big data might be thought to arise from its great scope, the vast number of data points included, and in the exclusion of human interpretation and bias in machine led data collection and analysis. This is sometimes described as 'data fundamentalism', 'the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth' (Crawford 2013). As such, it has received serious questioning. Nonetheless, even if it is to be ultimately rejected, it must be considered as a factor in understanding power relationships between data possessors and data subjects. Even putting aside the extreme data fundamentalist view of big data as truly objective and as such unquestionable, the thought that big data is very powerful and at least sometimes more accurate than other sources of information is one which critically shapes the ethical questions about the nature of the data subject and the concomitant power relationships with data possessors and users.

Secondly, big data may be thought be beyond the power of human beings to understand, and hence, again, perhaps beyond the capacity of human beings to question. This epistemological issue again places data subjects, as well as any users and producers of data, in a vulnerable position of relative powerlessness. Related to this, the replacement of understanding based on causation with correlations produced by crunching vast amounts of data again challenges ways in which 'ordinary' humans habitually think and comprehend their world (Mayer-Schonberger and Cukier 2013, pp. 50 ff.), perhaps again rendering them relatively powerless and lost in the face of what the data tells them.

Thirdly, somewhat in tension with the idea of data fundamentalism, there are concerns about inaccuracies in data and what might be called 'lossiness' (Busch 2014, p. 1735) from the often deliberate stripping away of context such as the original reasons for gathering data in a particular form (Mittelstadt and Floridi 2016, p. 322). This 'ontic occlusion' naturally again raises ethical issues for data subjects who may worry with good reason that they will be judged or manipulated unfairly on the basis of a poor understanding of their true characteristics. Coupled with a (false) idea that this data is actually objective and accurate, these worries deepen. Loss of the context of information will be a particular concern in this chapter.

Fourthly, these concerns combine in the fear that a data subject – an actual human being – might be lost entirely and replaced by their big data proxy: the profiled becomes the profile (Floridi 2012). A description of a person culled from numerous data sources, some known to the person, some unknown, arrived at via highly complex mechanisms and data analysis probably quite obscure to the subject, comes

to be treated as if it is the person themselves. The data profile then in a way becomes the person. This then leads to the problem that the data subject is actually distanced from interactions with those who use data – interactions are with the fictionalised data proxy and two-way communication between the real living data subject and those who use their data profile become perhaps impossible.

In thinking about this, it is important to consider how far and in what ways this differs from a more everyday case, because it is the apparent gulf between big data and the everyday situation which is the particular source of ethical challenge. Note that although each person has special or privileged access to some personal information, such as the exact way in which your shoes hurt when you walk, or whether you do find being tickled unpleasant despite laughing, it is of course often the case that others can know things about a person through a different route – for instance, someone else can spot that someone is limping – and sometimes a third party can know something that a person does not themselves know – from mundane things such as whether someone's shirt is hanging out of the back of their trousers, to more complex things like picking up body language and tone of voice to determine that although they say they admire someone, they are really ambivalent about them. The latter example is useful, because the person who works this out may do so without really knowing how, and hence this intuitive information gathering forms a rough proxy for the complexities of big data analysis. Now, although having such discoveries fed back to you by a third party can indeed be unsettling, in such an example, the possibility of communication between the parties mitigates the situation. It is thus perhaps the distance of the data subject from the whole process of data gathering, analysis, and use, and the consequent obliteration of lines of interaction, which renders the issue of substituting proxy for real subject most troubling.

This then leads nicely on to a fifth epistemological issue – the standing of communities of knowers in relation to big data. As soon as we recognise the importance of the community of knowers in questions of epistemology, questions of ethics are inevitable as the relationships between different knowers and potential knowers come into view. Philosophical discussions of testimony have made clear how those in different positions might legitimately have greater authority over certain claims to knowledge than others, and how in certain circumstances, the testimony of some may be legitimately doubted (Hume 1955). Yet, this can also lead to instances of testimonial injustice where certain speakers are unfairly not heard or not taken seriously (Fricker 2007). Where it is the individual themselves who is the subject of the disputed claims, then any injustice is magnified, and adds weight to the disempowerment of the subject.

We can see then how these issues affect fiduciary relationships (Mittelstadt and Floridi 2016, p. 328). Such relationships of trust, such as those between a doctor and a patient, are premised on many things and these include an unequal relationship of power and knowledge, where the doctor has knowledge that the patient lacks, about the practice of medicine in general, and about the patient themselves in particular, from test results and medical examination. In turn, the doctor must ideally be truthful and must listen carefully to the patient and gain an accurate account of them and

their case: there needs to be good communication and the subjective views of the patient are of great importance. Shoehorning this into a model of biomedical big data places all this in jeopardy, especially if the big data is seen to be unquestionable by even the doctor, and especially if the proxy is substituted for the actual patient. As Mittelstadt and Floridi say 'Traditional fiduciary "healing relationships" do not scale well to Big Data' (2016, p. 329). Later in this chapter, I shall consider how certain uses of data may disrupt such healing relationships to the point at which the power of the patient may potentially be increased rather than decreased.

3.1 Challenging Data Neutrality: An Examination of Some Key Common Epistemological Assumptions Concerning Big Data

It is also useful to flag some commonly made assumptions about the value of data, which again are sometimes stated explicitly, but are often simply implied. As we shall see later, looking at these head on will help to show how exactly ethical problems with biomedical big data may arise, and hence pave the way for addressing these problems.

The first culprit is the idea that at core, data itself is neutral – generally implying it is only by being used in various inappropriate ways that it can raise ethical questions. Not all would agree with this claim, but it is nonetheless subtly influential. For instance, Grady Booch writes, 'As an insider to computing, I know that data is morally neutral' (Booch 2013).

Data is not, and never can be, neutral for the simple reason that data always stands in some relation to some agent as knower or as potential knower of that data. If it doesn't, it's not data, it's some fact about the universe somewhere to which no human being yet has any relation. There is data collected in big data sets which nobody has ever scrutinised directly, which might have been gathered in an automated manner, yet it still has a potential link with human agency through the very fact of its existence in a data set. Data is a moral vector, reaching out morally in two directions, both to those (individuals and groups) who actually or potentially observe or manipulate the data; and to those (individuals and groups) whom the data is 'about' or who potentially may be affected in some way by data possession or use. The distinction between moral agents and moral patients may be applied here, with the former group being data agents, the latter data patients. (The latter may include other morally significant creatures as well as humans, but for practical purposes here, I am assuming for the sake of argument that only humans can be moral agents.) Since simple 'facts' about the universe themselves are also potentially 'about' individuals or groups, and since moral patients need not be moral agents, my points here about the lack of moral neutrality of data have taken as their starting point, the observation that as soon as a 'fact' becomes a point of data, this establishes a relationship of some kind to human agency. The relation of

the agent to an item of data may in many instances be morally innocuous, but it can never be neutral in the sense of being outside of the realm of ethical concern. This is by virtue of the very fact that it is contained within the domain of human awareness and of potential interaction with other agents. However, much more can be said to put flesh on the bones of this broad claim about the ethical status of data.

It is important to stress this seemingly abstract point since grasping it is crucial in order to understand the tensions between impersonal, algorithmic, technological machine driven notions of data, and personal concerns with information and communication – the link between big data and small talk, as the chapter title has it. Understanding these tensions is one key to understanding the ethical problems that arise with biomedical big data. To understand it, we must tease out certain aspects of what might be assumed by claims of data neutrality.

The notion of the essential neutrality of data makes perfect sense in a certain broadly construed context: that of the Enlightenment ideal of the pursuit of universal, timeless, scientific knowledge which is in principle accessible to all of humanity, in a universe which is essentially comprehensible, and which looks the same from any point of view. There is much to be said for such a picture and for the democratisation of access to knowledge which goes along with it. However, it is essentially an understanding of the ideals of knowledge stripped of any context, or rather, in which knowledge is seen *sub specie aeternitata* – from the point of view of the universe, an all-encompassing, all-seeing context. But this clashes with localised, personal and subjective norms of knowledge which come to the fore in certain other contexts – as with the case of personal, including biomedical, knowledge. This will be explored in greater depth when we turn to look at social research on the communication of such information.

The neutrality of data also often comes with the implied assumption of what I call 'data eternity' or 'data timelessness' – which is not of course to say that data has always existed, but that, in examining the value issues involved, there are no essential issues concerning timing. This is often expressed in the notion that sharing data has no costs – that there are no rivals in information and so we might as well share with whoever wants the data, it makes no difference, other things being equal. 'Information is what economists call a "non-rivalrous" good: one person's use of it does not impede another's' (Mayer-Schonberger and Cukier 2013, p. 101). What is being eclipsed from such views is the relation of the data to the agents who has access to it, as well as to those whom the data itself concerns. This relationship comes into being at particular times and in particular contexts and hence again essentially involves at least potential ethical issues. To understand and include notions of agency is essentially to make reference to action and meaning, neither of which makes sense without reference to notions of time. Again the assumption of data eternity fits well with a scientific view of the universe as abiding by unchanging laws and as looking the same from every direction. It just fits terribly badly with how particular human agents view knowledge in their everyday social lives. Try telling a journalist that sharing data has no costs or that timing does not matter. Any school child will also tell you that the kid who is the first with the news that the headmaster

has emptied the school bank account and run off with the French teacher will see their social status shoot up. Such issues about information and timing will also be explored in detail later.

Another aspect of data neutrality which also helps to eclipse the relationship of data to the data knower is what might be called the 'billiard ball in the cup' view of data and data sharing, or, more briefly, 'data atomism'. This is quite simply the idea that data is the same whoever possesses it and from whatever point of view: that in being shared, it is simply copied as in some ideally perfect photocopy machine, and in being transmitted from one point to another, from one person to another, what is transmitted is always exactly the same, and, importantly, the container or repository of the information is also unchanged by this. Now, obviously, data may be corrupted, degraded and lost by the process of transmission. But this is understood as a departure from an attainable ideal. In looking at the findings of some social science research on the communication of medical information, this ideal will be questioned in ways which aim to shed further light on the ethical tensions raised by biomedical big data. It will be particularly important critically to consider the underlying assumption that the repository of the information – in this case, the particular agent – remains the same, and indeed, to question the assumption that what is transmitted – the data – is likely to be the same. In the context of the transmission of personal information, these assumptions will be radically questioned, and it is in their questioning that the key ethical issues arise.

4 How the Argument Will Proceed from Here: Further Details of Methodology

As discussed, the aim in this chapter is to look widely at possible ways of thinking about the ethical issues of biomedical big data. As such, it should be seen as starting to explore the issues, rather than as aiming to produce polished answers, although suggestions for directions forward will be made.

This chapter draws upon qualitative social science. Despite the very power of big data, it has been remarked that in exploring the social and ethical issues it raises, we need more than the methods of big data itself. Kate Crawford expresses this point when she says '... data scientists should take a page from social scientists, who have a long history of asking where the data they're working with comes from, what methods were used to gather and analyse it, and what cognitive biases they might bring to its interpretation,' and argues for the need to use rigorous qualitative research (Crawford 2013). Social research, and in particular qualitative social research, will be useful not just for its empirical findings but for its rigorous and long-standing attention to conceptual and epistemological questions in the gathering, interpretation, and use of data (see e.g., Hammersley and Atkinson 1986), and not least, in its critical empirical and conceptual probing of the nature of the individual subject and in social relations.

The chapter also draws on debates within genomics research. It has already been pointed out that ethical issues in genomics overlap considerably with questions about biomedical big data, and have an advantage that they have been discussed for slightly longer. The points of overlap refer not just to the sheer size of the data sets in genomics; the way in which genomics research has pioneered data sharing and open access to data (International Strategy Meeting on Human Genome Sequencing 1997, 1996; Wellcome Trust 2003; Toronto International Data Release Authors 2009); and the abilities to aggregate data sets with others, sometimes unleashing new and unexpected data. Another point of commonality, which may not at first stand out but which is worth stressing, is the theme of speed. Talk of big data and its power, promise, and perils ubiquitously stresses how quickly big data is being gathered and analysed. Likewise, it has also been often observed that genomics research is frequently accompanied by hype, not just about its potential, but about the speed at which progress is being made (Boddington 2012, p. 77). This hum of speed will prove to be an important background noise to notice in understanding ethical issues that arise concerning the use of information.

An additional and vital point of overlap concerns the imagining of the human subject, and the sensitive nature of the information. It has already been pointed out that ethical issues arise in biomedical big data in part because of the sensitivity of the subject matter to the individual, coupled with the great usefulness and potential of the data, which creates tensions between values of the private and the public. And some of the most pressing issues concerning the ethics of big data involve how human subjects are described, analysed and manipulated by big data, to the point where questions arise about the very nature of the data subject. And genomics, likewise, raises questions about the nature of the human subject. Even though it is important to dismiss the claims of genetic essentialism that there is nothing more to the human being than a genetic profile, nonetheless it is not possible to understand the ethical issues around the handing of genetic and genomic data without understanding the strong link of DNA to an individual's identity and claimed uniqueness, including the imagined constructions of such links. And yet at the same time, many of the ethical quandaries come about because of how genomic data link us to others – to biologically related individuals, to population groups, to our ancestors. At once highly personal, some of the information about our DNA just is information about others. For instance, the information that a father carries a gene on his X chromosome lays open to view the information that his daughters also carry that gene. So there is a real sense in which genetic information about one person may also be claimed by others. Additionally, there is a sense in which human genetic information in general may be claimed by humanity in general, precisely because of its subject matter – as expressed in the dictum 'the proper study of man is mankind' (Pope 1796) and as illustrated in the fanfare around the announcement of the sequencing of the human genome at the White House in 2000, hailed by Clinton with the words 'Without a doubt, this is the most important, most wondrous map ever produced by humankind' (White House 2000). It is essential to our experience of ethical issues that these frequently and centrally concern the balance between how individuals are valued and how the wider social group, however defined, is valued.

Hence if big data presents ethical issues concerning the nature of the subject and how subjects relate to others, then these issues are also archetypally laid bare in considering genomics.

This then leads on to an explanation of the attention in this chapter to some ethical issues in clinical genetics, and in particular communication of genetic information, both within the family and also from professionals to patients. This is not 'big data' but nonetheless, is essentially concerned with information which inhabits at once a private, personal realm, and a realm shared with biologically related others. Tensions thus exist about communicating this individual-yet-shared information. Some of these explicitly arise in tension with a model of ethics which prioritises the sharing of information. Moreover, it is information which is at once iconically 'about' an individual, yet which is passed on to that individual by someone else as a result of technology and data analysis. The individual who passes this information on initially is usually a professional, and usually in a context where that professional has far more understanding of that information than the subject of that data. This again reproduces some of the feature of big data – even to the extent that such is the speed and power of advances in genomics knowledge, that information passed on at one time may later be found to have unexpected significance beyond current understandings of even competent professionals.

Hence, putting these issues together with the desirability of using the methods of social science research to shed light on the ethical issues of big data, this chapter explores qualitative social research into the communication of information within clinical genetics as one promising method of gathering small-scale insights to illuminate questions about the sharing of biomedical big data.

The chapter also looks briefly at ethical issues concerning the use of e-health (healthcare practices supported by electronic processes and communication) and m-health (healthcare and public practices supported by mobile devices) technologies in developing countries. One set of issues raised by biomedical big data concerns big data divides, and more generally, issues of disempowerment of data subjects. Perhaps related to this are issues of how the disruptive power of big data may lead to changing fiduciary relationships. E-health and m-health technologies are already using data in novel ways and hence, again, raise some of the same questions as those raised more generally by big data. In addition, a particular reason exists for looking at their use within developing countries. If you are concerned with possible disempowerment, study the disempowered.

5 Genetic Information, Communication and Ethics

Now we turn to look at some findings from qualitative social science research about personal and family communication concerning genetics. This will of necessity involve making broad suggestions. It should also be noted that this section draws upon a large body of research from which certain key conclusions are condensed for present purposes. To recap, the aims of this are to explore differences in

norms concerning such personal communication from norms of knowledge more appropriate to the sharing of big data, in order the better to understand the nature of the ethical issues that could arise concerning big biomedical data, and to begin to look for ways such ethical issues might be addressed. We are looking at a picture of data transmission which might work well in some contexts – for instance, in some areas of big data, perhaps those not involving human subjects directly – but not in others. The context of the information, how it is gathered, whom it is about, will all turn out to be crucial. Genetics is chosen for examination because of how it highlights and inhabits the juncture between ideals of sharing impersonal scientific knowledge – or knowledge which somehow belongs to all of humankind, given that it stands in some relation to a notion of humankind's essence – and at information which is at its most, even uniquely, personal. Indeed, some of the research I will look at explicitly concerns the ethical quandaries that arise from the clash of these public/private views of knowledge. Much of this analysis explicitly questions the assumptions of data neutrality gestured at earlier. I shall firstly outline some key findings before discussing their significance for how we might think about the ethics of biomedical big data.

Firstly, information which is transmitted in a medical or clinical context may be understood very differently by professionals and by patients, yet 'deficit' models of comprehension appear inadequate to capture this phenomenon. A large body of research shows that there are often very great differences in lay and in professional understanding of medical and genetic information (Walter et al. 2004). Yet, it is unwise to attribute this simplistically to a 'deficit' in lay comprehension. For example, it is clear that these two domains of knowledge may be considered to have different qualities, and that it is wrong to assume that knowledge formally acquired from a medical professional is necessarily superior, or represents an ideal to which the knowledge derived from more grounded experience must aspire (Gregory et al. 2007). Sociologists have uncovered quite surprising misunderstandings of basic genetics amongst patients who, crucially, nonetheless display a good understanding of certain key aspects of how to manage their medical condition, and lay understandings of patterns of inheritance may be quite at variance with 'scientific' understandings (Featherstone et al. 2005). What is clear from much research is that a simplistic model that the medical professional takes a bit of information out of his or her brain and somehow gets that very same bit of information into the brain of a patient is very wide off the mark.

Understanding the medical information depends crucially upon the personal context and experience, including past experience and what might be called 'readiness' to absorb information. For instance, it has consistently been found that comprehension of patterns of inheritance and risk are better in cases where individuals have direct experience of a condition within their immediate or extended family. Practical understanding of a disease is gathered through experience of watching a relative manage a condition, and this translates to better understanding of the more abstract notion of one's own risk for that disease. The knowledge is then consolidated through practice; mere transmission in the clinic from a medical professional may not be enough for the knowledge to 'settle into place' with

its implications and import recognised and felt (Gregory et al. 2007). Research has found that attitudes towards genetic diseases varies widely, depending on factors such as mode of transmission, family history, prospect for cure, hope of prospect for cure, the nature of the condition, age of onset, and so on (Featherstone et al. 2005). However, professionals are often blithely unaware of how discordant patient understanding is: the patterns of understanding and misunderstanding do not map neatly with what the medical professionals think they have communicated, nor with 'scientific' understandings; one reason for this may be communication of information within the relatively enclosed setting of the clinic and outside the context of everyday experience (Boddington and Gregory 2008a).

Crucially, individuals can actively resist knowledge. W.K. Clifford argued in his paper 'The ethics of belief' for certain norms of knowledge, including the claim that we ought to found our beliefs on what evidence is available (Clifford 1877). Starting from William James' response in 'The will to believe' (James 1896), philosophers have debated what norms should govern belief formation and preservation. But whatever the norms of belief should be, empirical research shows that individuals frequently depart from Clifford's strict ethics of belief. Knowledge may be refused or avoided because of its wider significance, significance which another person may be unable to gauge, and which may relate to a person's sense of identity. To illustrate, one young woman discovered she was a carrier for haemophilia, inherited from her father, at the same time as discovering that the man she thought was her biological father was in fact her step-father. She then refused to learn more about haemophilia or investigate further what her carrier status meant for her. Interviews revealed that to her, this represented a kind of disturbance to her sense of integrity and relationship to her (much loved) step-father. The knowledge of her carrier status concerned not just an 'objective' fact, but had personal import and meaning that spoke centrally to her biographical and personal history and relationships (Boddington and Gregory 2008b). The information changed her in that it shaped who she felt she was and how she related to important others. Only on a certain view of the ethics of belief and of the value of information perhaps coupled with an imperative of 'healthism' (Crawford 1980) could this be seen as an 'irrational' or unethical refusal of knowledge.

The importance of the readiness to receive and understand knowledge means that we must pay particular attention to timing, and this is key both in the reception of information and also in judgements about the transmission of information. For example, in one study, teenage girls had been marked by professionals as having given informed consent to carrier status testing; yet when interviewed later as young adults, many claimed to have had little or no idea of what the testing was about. The researchers concluded that as teenagers, the girls were not ready to receive or understand the significance of information about their future reproductive risk. The information was passed on, but lay inert, as it were, until a stage of life had been reached in which its significance could be grasped (Boddington and Gregory 2008b). Knowledge may be 'forgotten' then recalled, and perhaps then seen

differently, at different life stages as appropriate. The knowledge then is not 'passive data' stored mindlessly as if coded in a machine, but inhabits a world of personal meaning.

Likewise, certainly within the family and with the transmission of personal information, disclosure is rarely a single incident, but takes place over time, with loose boundaries around the transmission of information, and understanding is often not reciprocal between parties (Arribas-Ayllon et al. 2011). Timing of disclosures may be prompted by external events or may be made with careful judgement about the 'right time to tell'.

Notions of timing and of 'readiness to receive' information are recognised by those who are considering passing knowledge on, and it is notable that these judgements frequently make reference to character. These may be judgements about who deserves to know; who is able to 'cope' with knowledge; who is placed in the right position within the family to be told; as much as about who needs to know. Someone judged to be of general 'poor' character may not be trusted to be able to 'handle' information, including not being trusted to pass it on appropriately to others (Boddington and Gregory 2008b; Arribas-Ayllon et al. 2011). Judgements about communication may depend not just about character, but also upon location in the family, and the division of labour in transmitting information may be shared: studies have sometimes found that it tended to be males who decided who ought to be told genetic information, but females who did the telling (Arribas-Ayllon et al. 2011; D'Agincourt-Canning 2001). This is very possibly culturally specific, but that observation only underlines further the point that context is crucial in understanding how these practical ethical judgements are made.

5.1 Ethical Issues in the Personal Communication of Genetic Information

One reason for exploring the communication of genetic information is that ethical questions have arisen which mirror closely some ethical questions concerning biomedical big data. A debate has been occurring for some time about the rights and responsibilities of the transmission of genetic information. On the one hand, an information- and autonomy-driven model of medical and genetic information has it that, not only is such information of great value in pushing forward research and medical progress, but also that individuals have a right to access such information about themselves (See e.g., Montgomery 2003). On the other hand, precisely because it is personal information, relating closely to notions of the self and identity as well as to medical information which regarded as private, individuals also are accredited with rights to control that access. This gives rise to a host of problems, especially concerning rights and wrongs of transmission of such information within the family (see e.g., Skene 1998). Does a patient have a duty to pass on personal genetic information to other family members, if this might be

of relevance to their relatives' own disease or risk status? Or does a patient have a right to keep such information confidential? The implied virtue of obtaining relevant genetic information about oneself can be inferred from the debates about 'genetic ignorance', which wording hints at a deficit model in those who lack or refuse such knowledge.

This emphasis on the importance of access to genetic information has been stressed by developments in professional practice regarding genetic counselling as the provision of information to the enhancement of patient autonomy (Kenen 1984). Further drivers to the emphasis on the importance of genetic information can be found in considering notions of geneticisation, a phenomenon whereby genetic explanations and accounting of health and disease come into ascendancy (Lippman 1991, 1992; Dreyfuss and Helkin 1992). Indeed, the UK General Medical Council has issued guidance which urges doctors to encourage patients to pass on such information, and allows the possibility to circumnavigate usual rules about patient confidentiality in order to facilitate giving potentially relevant genetic information to relatives if patients demur (General Medical Council 2009). This implies that the value of patient autonomy, privacy and control over their personal information is being squeezed by the value of the transmission of that information to someone else who might need it, and it can be argued that in drawing up these new guidelines the GMC is diluting the norm of individual patient confidentiality (Boddington 2010). Hence, this models a key question in the ethics of biomedical big data – how far is sharing this, in the name of the wider public good, to trump individual control over personal data?

Despite encouragement to pass information on, patients often fail to do so, often even after indicating to professionals that they will. So are these patients 'ethically deficient'? Studies and analysis of family communication suggests strongly otherwise. What is deficient is the ethical framework within which such responses are understood and found wanting.

Far from acting in an ethically irresponsible or selfish manner, researchers found that individuals often considered very carefully how to communicate (Arribas-Allyon et al. 2011). The framework of rights and duties in which the debate has been drawn up in academic literature does not seem to capture what most concerns people in practice. Amongst respondents interviewed, there is little or no talk of 'rights' of access to information, but concerns are expressed, as indicated by the discussion above, about who is a good person to tell, and who is the right person to be told; concerns are not so much what to tell, as who to tell, when to tell, and how to tell. These may be quite localised and nuanced judgements, for example frequently relating to local knowledge about a person's relationship status – for instance judging that there is no point in telling a teenager she might be a carrier for a condition if she 'hasn't even got a boyfriend yet'. These concerns all reinforce the ways in which the information – the data to be passed on – is not 'neutral' and timeless, and that it will have various meanings for the different people involved (Boddington and Gregory 2008a, b; Arribas-Ayllon et al. 2011).

An interesting finding concerns how information itself shapes notions of ethical responsibility. From an 'objective' perspective, notions of genetic relatedness to a

proband might determine responsibilities to communicate genetic risk information to that other person. However, this is premised upon a biological 'objective' notion of the family. In contrast, researchers found that the very fact of presence or absence of communication with an individual itself acted to form notions of relatedness, which in turn then were drawn upon in justifications of patterns of communication. In other words, the very fact that a biological relative had not been spoken to for some years, was used to construct this person as falling outside the realm of the family and hence, outside of the realm of responsibility to communicate (Arribas-Allyon et al. 2011). There is no set realm of relatedness which gives duties to communicate: rather, patterns of communication themselves serve to mark out the territory of ethical obligation.

Importantly, it is not simply that an ethical imperative to communicate information was replaced with more nuanced judgements about when, who, and how to tell: often, it was supplemented by taking on additional moral responsibilities. This was especially for those who were not particularly close, where communication might be more difficult. Frequently the task of surveillance for signs of disease was undertaken to judge if and when it would be a good idea to disclose information (Arribas-Ayllon et al. 2011). For instance, it might be judged that there would be no point in disclosure if it turned out that one branch of the family was not affected. It is to be noted that this cannot be seen as simple moral laziness, since not only were such judgements based upon concerns that simple disclosure might not be the best thing to do, but also such strategies involved long term and careful practical ethical work.

So, we can conclude that not only do the professional ethics of the responsible communication of information fail to capture the practical ethics of those involved, but that individuals are able to resist this professional information ethic by proposing their own strategies of observation and surveillance. One could add in addition that the strategies of surveillance and waiting are still motivated by concern for health outcomes.

There are many different ways in which such responses to the perceived obligation to disclose, such as the adoption of surveillance strategies, might be understood and interpreted. It should be noted that the participants in these studies recognised implicitly that they lived in a moral order where disclosure 'should' take place (Arribas-Ayllon et al. 2011, p. 19). Arribas-Allyon et al. briefly draw on Deleuze's account of Spinoza's ethics in discussing their findings, to explain how non-disclosure may be understood without invoking communicative incompetence or a moral deficit (Deleuze 1988). Broadly, this provides a model for understanding that an individual is not simply an unaltered recipient of knowledge, but can be materially altered by understandings of genetic risk, giving a 'dynamic and material understanding of how individuals are not just positioned by a discourse of genetic knowledge but are 'modulated' (literally modified and intensified) by the idea of genetic risk. Thus, the modes of genetic knowledge (that is, affections, images or ideas) produce a powerful embodied imaginary of fatalism, which in turn resonate with ambivalence, self-blame and dread' (Arribas-Ayllon et al. 2011, pp 19–21).

In other words, the failure to communicate need not represent either a failure of epistemic ideals or of moral imperatives: it may indeed represent the fulfilment of these. An epistemic ideal which emanates from model of timeless, universal, objective knowledge, to be transmitted whole and unchanged from one separate, equal, equally placed, individual, who will remain essentially unaltered by the receipt of the knowledge, and who needs this knowledge to realize and to further their own autonomy, is misleadingly inappropriate in this context. In this context it may be neither an attractive epistemic ideal, nor an attractive moral ideal. 'The practical ethics of everyday experience is as much about waiting, watching, ignoring or forgetting; it has its own silent, endogenous logic that, despite the consequences, is resistant or oblivious to contemporary risk politics' (op. cit., p. 21).

For the purposes, it matters less how the strategies of resistance to a dominant ethic of information and communication are to be understood and interpreted, than to note their existence. It is precisely the point of the present discussion that there are context-dependent and local responses, so it is likely that in different contexts, different responses will be made.

It is interesting to note that there are some parallels in attitudes to the ethics of sharing data among groups of scientific researchers. Willingness to share may partly depend upon whether another party is themselves judged to be a 'good sharer', and this may involve not just judgements about whether they will adhere to ethical standards of curation and use of data, but whether they will, in turn and under the right safeguards, also share data. This may present ethical problems for instance, in the case of those whose reluctance to share may be based upon reasonable considerations, if these are not acknowledged as such. For example, sharing may have resource costs, and hence excluding someone as a poor sharer may disproportionately exclude those with fewer resources. Likewise, reluctance to share may be based upon previous poor experience of data exploitation, and hence, judging that someone is a poor sharer may bias against those who have also suffered in the past, hence potentially reinforcing injustice. To avoid such injustice, great care must be exercised before parties are excluded from sharing agreements (Boddington 2012, p. 198 ff).

5.2 Summary and Implications for Big Biomedical Data

It will be useful to recap this discussion and consider how the picture of information and communication that emerges contrasts with the picture emerging from biomedical big data.

One picture that emerges from big data can be roughly characterised as stripping away context; as favouring open sharing; as seeing no inherent costs to sharing; as prioritising speed, yet as being insensitive to issues of timing; and (although data is analysed and aggregated in ways which may often lead to surprising and powerful results, and although data may deteriorate), as understanding the data itself as remaining intact in transmission. In the particular case of a rational, scientific,

public health view of the sharing of genetic and other medical information, although there is a recognition of privacy and individual autonomy, this comes into conflict with the imperative to share, and there may be possibly increasing encouragements to prioritise public sharing over personal privacy.

In the case of personal communication of information, however, context is everything. Sharing is not open, but subject to local, contextualised and nuanced judgements which make essential reference to character and normative issues; timing is of the essence; what is transmitted is modified in its transmission, and those who transmit and receive the information, and their relationships with each other, may be changed by that very information. An ethical framework which pitches public open sharing of knowledge against privacy and rights may be resisted and even bypassed with alternative practical ethical strategies.

5.3 Information and the Resisting Subject

As has been noted, a significant ethical concern with big data is the reduction of the individual to a profile, of interactions with a human being replaced by the manipulation of a more or less accurate fiction culled from an aggregate of information. In the face of being reduced to a mere data proxy by organisations which have access to untold amounts of data, the human subject may become disempowered. As suggested above, this concern can be seen mirrored in concerns about geneticisation and the logic of healthism, and the corresponding ethic of personal responsibility to seek out and act on genetic information, which may skim over the far more nuanced reality of an individual's personal and social situation.

The social science research discussed above discovered ways in which people are resisting and bypassing these caricatures of their situation. In discussing this, Arribas-Allyon et al. (2011) consider how Novas and Rose argue that the fears about the impact upon subjects from geneticisation are exaggerated, and suggest rather that 'Genetic risk does not imply resignation in the face of an implacable biological destiny; it induces new and active relations to oneself and one's future.' (Novas and Rose 2000, p. 485). Using the notion of biological citizenship or biocitizenship (Rose 2007), new forms of social action arise, together with new social configurations and alliances which may place individuals within a 'genetic network' (Armstrong et al. 1998) and which involve new forms of 'biosociality' (Rabinow 1996; Taussig et al. 2003). That is, rather than being merely constrained by 'reduction' to a genetically determined subjectivity, this very identity is used to mobilise and to act. Conversely, some have argued that others may act and mobilise around resistance to the notion of biological or genetic identity (Plows and Boddington 2006). This is a complex area of research and discussion which there is insufficient space to examine fully here. Nonetheless it suggests an avenue of exploration for concerns about how biomedical big data may view and constrain the subject.

Hence examination of such notions of biocitizenship may prove extremely fruitful for future work on the ethics of biomedical big data. It must be noted, however, that this practical and ethical work of managing, exploiting and resisting identities constructed via the advances of scientific understanding, technology and information, is largely taking place outside the proscribed realm of formal ethical regulation, which has to inhabit relatively formal, institutionalised, professionalised and public spaces.

5.4 The Importance of Context: Building on Nissenbaum's Contextual Integrity

Mittelstadt and Floridi draw attention to the importance of context (2016) and likewise, the above discussion makes amply clear just how important this is to understand ethical issues in communication. Nissenbaum's work on contextual integrity in privacy (Nissenbaum 2004) provides a framework for looking at context around informational norms which could be expanded in ways making it yet more useful as a framework for understanding how ethical issues arise in biomedical big data and how they might be approached. Although Nissenbaum focuses specifically on privacy, the informational norms she discusses can be seen more generally as norms concerning communication, especially norms of personal information.

Nissenbaum argues that people are concerned that information flows appropriately (and indeed this is a good rough characterisation of the social science research findings drawn upon above). She concludes that we do not have a right to control personal information as such, but rather a right to live in a world where our expectations about the flow of personal information are, for the most part, met. These expectations are met by force of habit and convention and accord to key organising principles of social life. Contextual integrity is the 'harmonious balance of social norms with both local and general values, ends and purposes' (Nissenbaum 2009, p. 231). Protests against technology-based information systems can be traced to breaches of context-relative informational norms.

These informational norms concern senders, recipients, and information subjects. Among the critical variables are actors' roles, and the type of information involved. The norms can vary between contexts in terms of how restrictive, specific and complete they are. Nissenbaum acknowledges how in fields outside law, such as philosophy, there are some richer accounts of the nature and importance of privacy. For instance, she mentions discussion of the importance of privacy and informational norms in shaping the nature of our relationships (Rachels 1975), and indeed Rachels' point is echoed in the social science research findings that lines of communication shape who counts as family which in turn shapes notions of responsibility to communicate. Even in different public settings, individuals can act according to different norms of privacy and norms of informational flow may be open and fluid in certain contexts such as those of friendship.

Nissenbaum argues convincingly that her approach recognises a richer and more comprehensive set of parameters than other approaches to privacy. I suggest that these can be made richer and more nuanced still if the lessons from social science research are incorporated. The compass of her framework means that it may be hard to apply without detailed understanding of each context, and it is here that social research would be highly valuable. It is important to recognise that this may often mean that there is no easily accessible answer to what a particular context requires. The 'traditional' framework of rights and responsibilities and the opposition of public and private information may be inadequate to capture the contextual informational concerns of all actors as we have seen, and indeed finding out what these contextual norms are may be itself a complex and delicate process. Moreover, the norms may vary from individual to individual, depending on their place in the social fabric and in individual characteristics; indeed, an individual's very location within a network of social relations may be determined by informational norms rather than vice versa, as we have discussed above.

Nissenbaum's account could be supplemented with further variables: it is not simply the type of information which is important, but the manner of the information transmission – how a person is told – and the timing. Additionally, a rebuttal of any simple idea that it is the exact same parcel of information which is transmitted from person to person is needed. Determining which actors have which roles in communication may be complex and may depend on highly localised knowledge, and roles may be divided and dissipated throughout a community. To the variable of actors' roles, must be added the insight that an actor's role may vary by social location and by character. Attention to the detailed norms of actors' roles may help to protect such actors from the potentially disempowering use of biomedical big data.

Nissenbaum recognises the complexity of informational norms and that they may be multidimensional. The notion that there is a final answer to the contextual norms of a situation then unfortunately also needs to be questioned. There may be no one 'right' answer. Now, in a legal context, or a context of institutional regulation of ethics, to have more than one answer to the question of what contextual norms information to follow may be a disaster. But within a practical context, this may simply reflect the reality of an ethical dilemma. Nonetheless, within such dilemmas, individuals can perhaps find ways of ameliorating or circumventing the situations, for instance see the 'surveillance' strategy described above.

Nissenbaum recognises the problem that the contextual integrity approach might be inherently conservative in looking at current norms. This indeed would be a problem for dealing with developments in big data. But what is useful about the notion of contextual integrity is that it alerts us where to look in forming new regulations and new ethical responses. A conservatism might arise if attention to contextual integrity was coupled with an urgency to find a solution *right now*, since that would pose the danger of freezing norms – or of producing a volatile solution since we may well find a mass of conflicting norms. A conservative response might also lead to the problem of over-restrictive regulation (Mittelstadt and Floridi 2016, p. 306). Where laws are concerned, current cases must obviously be given answers,

but care must be taken not to set undue or hasty precedent in ethical regulation in such rapidly evolving areas. This is especially the case given that individuals and society needs time to adjust to the challenges of biomedical big data. There seems reason to suppose and to hope that such adjustments may be made by individuals and groups coming to terms with and exploiting biomedical big data, as much as by individuals and groups resisting reduction to their data proxies. A slow, careful, localised, observant ethic is needed to cope with the rapid all-consuming advance of big data.

6 E-Health and M-Health Initiatives, Data, and the Subject

The preceding discussion suggests that there will be many different contexts that we need to examine to consider norms of information, and that lines of communication themselves shape relationships and hence shape norms. From this it is implied that changing lines of communication will be likely to shape and change the resulting norms, as well as the fiduciary and power relations between those involved. Such changes can also be seen in a different area, that of the use of e-health and m-health technologies especially in developing countries. These essentially involve the use of data and rapidly changing patterns of access to data. E-health and m-health technologies are rapidly transforming access to health care, and especially having an impact in less developed countries (World Health Organisation 2011, 2015; Future Health Systems; Blaya 2010). Social networks and social media may also be used in health care (Coiera 2013). This data may be used in ways which distinguish it from big data, especially as applied and used directly by patients. Nonetheless, it is a rapidly expanding use of large amounts of data and as such, worth examining briefly and in outline for possible clues as to the ethical challenges of big data, its consequences in different contexts, and how to respond. Of particular interest may be instances where there is 'leakage' of data from one context to another, as discussed below.

E-health and m-health technologies often make use of large amounts of shared data. Devices feed data back to patients and local health care providers. Although there are important ethical questions to be asked (such as who 'owns' or controls the data, since this is often now in the hands of mobile phone companies rather than managed within a standard health care setting), there are reasons for optimism regarding these technologies. These technologies can empower individuals and communities, for instance in helping to educate local health workers and to transform individuals into effective informal health workers. The individual, especially one who is managing a chronic condition, (an increasing problem in developing countries as well as in the developed (Olmen et al. 2011)) may become an 'expert patient' (Decroo et al. 2012) who knows more about their condition than their health care professional, both in terms of their own case, and in terms of general access to information concerning their disease (Keilman and Cataldo 2010; Wasson et al. 2012). Some of this information regarding diseases and treatments may indeed have

been arrived at via use of social platforms and via biomedical big data. Nonetheless, the way in which these technologies generate vast amounts of further data in their use, gathered remotely by third parties, can be a genuine cause for concern.

The traditional fiduciary relationships between patient and health care provider are unsettled, but not always in a negative way (Wasson et al. 2012; Alpay et al. 2011); the collection and analysis of data and the use of mobile and electronic technologies in the hands of millions of patients across the globe has the potential dramatically to change the nature of the medical encounter and with that, to reorient the whole setting within which our current broad conception of medical ethics has been framed (Boddington 2013). The standard professional relationships and the ethical codes that have arisen around them are premised upon an unequal power relationship, including inequality in terms of knowledge. Codes of ethics aim to iron out these inequalities and to protect the vulnerable patient from the consequences. But the very use of technologies can change the power balance involved and in particular, the epistemological imbalances. Technology is reshaping and undermining the very notion of a profession as we knew it.

Our current codes of medical ethics have also been framed within the relatively closed space of the 'clinical encounter', a space whose borders are challenged dramatically by e-health and m-health practices. There are pluses and minuses. A danger presents that ethical 'leakage' from under the umbrella of codes of good practice may occur. For example, health information about individuals is now in the hands of technology companies and commercial organisations which do not abide by the same codes of medical ethics, such that the health information is no longer in the safe harbour of a particular profession operating under certain codes of conduct with concomitant penalties for abuse. Hence, in one context, data may seem comprehensible and manageable as individual health data; in a different context, that same data may be part of a far more messy big data setting where its significance may extend far beyond the health sphere. On the plus side, patients can take control of their health, a benefit that is all the more important in settings where access to healthcare and professional services is otherwise difficult or non-existent. As opposed to the relatively simple model of communication within the clinical encounter upon which medical codes of confidentiality and informed consent have arisen, communication may be less direct, it may involve interpretation by third parties locally, and knowledge may be dissipated throughout the community and more widely through the use of social media. Democratisation of medical knowledge may occur when information is spread and shared in local communities. Hence, this can raise hopes that the imaginative and resourceful use of e-health and m-health technologies may be used to breach the big data divide, and could be used effectively to address inequalities. With increased knowledge about their conditions, patients and communities will be in a better position to communicate their concerns and needs. Indeed, where the worst inequalities are, that is where the biggest add-on value may be found, and where the impetus to make use of such technologies is the greatest (Olmen et al. 2011).

Lessons for big data ethics can be learnt too from the lessons that are being learnt about obstacles to benefiting from e-health and m-health technologies.

Health literacy is vital to the effective utilisation of health technologies and is a complex concept, often understood in different ways (Nutbeam 2008; Peerson 2009). However the technologies themselves can play a role in enhancing health literacy. Likewise, we need careful consideration of the literacies needed to engage in an empowered way with big data more generally. It should be noted then, that although such literacies will be complex, the literacy which individuals and communities need to benefit from, interact, and effectively control information and technology, and which will make for expert patients and empowered communities, does not necessarily need to be the same as the literacy possessed by professional 'experts'. It is also important to note how localised and embedded in particular cultural contexts ways of addressing informational disparities may be (Leonard et al. 2013).

It should also be noted that there can be very surprising differences in the effectiveness and implementation of remote technologies for managing health, and that what works in one context and one country may not work in another. There is no escaping the need for grounded local knowledge. For example, in one country there may be widespread ownership of mobile phones, but very patchy signal; in another country, good provision of health care buildings but these are badly equipped or short staffed; in another country, good mobile phone coverage, good signal, but poor rates of literacy (Leonard et al. 2013; Ahmed et al. 2014a, b). These findings can only reinforce the need to look in very close and contextualised detail at the impacts of big data.

7 Concluding Remarks

This chapter has aimed to mine various sources for insights into the ethical challenges of biomedical big data. This has of necessity added even more complexity to the issue, especially given the need to examine local context to understand the ethical cost of the attrition of context that may occur in big data. One conclusion must be that it will be necessary to continue to cast our net widely in looking for creative and human responses to the challenges of biomedical big data. Hence, here this chapter can make no more than some general comments and give some broad indications.

Just as the dangers of big data are partly those of supposing that we can systematise everything, that this can start from a neutral basis and, via a machine led process, end on a neutral basis of 'objective' truths and insights, so too there is a danger that a response to the ethical challenges of big data must be a (purely) systematised ethics. Although we surely need ethical regulation and law, and although we surely need systematic thinking in ethics, to consider that this is the whole of ethics is just as bad a mistake, and, indeed, the same kind of mistake, as overplaying the benefits of big data and of downplaying its shortcomings. Wittgenstein once wrote in a letter to a friend, 'It is plain, isn't it, that when a man wants, as it were, to invent a machine for becoming decent, such a man has

no faith' (Wittgenstein 1967, p. 10). If we think of over-reliance upon the ability of regulatory machinery to solve the ethical challenges of big data as a placing too much trust in a 'machine for becoming decent', we can take the warning that we need to think outside of the regulatory box as well as inside it.

The research on family communication about genetic conditions found that knowledge is dissipated and spread throughout the family and across time, with understanding of information variable across time and context even within one individual. Likewise, the ethical workload is divided up between individuals, spread out across time, and 'official' accounts of the ethical workload are challenged. New uses of biomedical data and mobile devices are taking the traditional practice of medicine outside the relatively closed setting within which ethical and regulatory frameworks have arisen and function. This precisely indicates that ethical responses must also be wide-angled.

The broader social and political setting of any regulatory framework is a vital component. If, for example, the models of various responses of biosociality and biocitizenship might suggest that similar socialities might arise more broadly in response to big data – if 'bigdatacitizenship' is not too ugly a term – then the freedoms and resources of civil society and open debate are needed. Recognition of individual citizens or groups and reciprocal processes of communication by big data custodians, analysts and users may be, realistically, hard to assure by regulation, but it can surely be addressed and encouraged. This is not to downplay the complexity of potential issues. For example, many groups organised around lobbying for patients receive funding from pharmaceutical companies who stand to profit from the uptake of biomedical and pharmaceutical responses to conditions, which raises potentially troubling questions about power bias and the exclusion of other viewpoints (Plows and Boddington 2006).

We saw too that the apparent breach of expected ethical standards by patients and family members who hesitated to communicate genetic information was actually motivated out of concern for future health and wellbeing, albeit not in the precise way as expected by professionals. This then might form a clue to ways of addressing fears about biomedical big data. If the usage of biomedical big data coheres with patient and public expectation of health gain, this might help to foster trust in biomedical big data. Encouraging health literacy and health education would be beneficial. However, this will be a complex task, one which there is insufficient space to examine in great detail here. It is wrong to think that there is only one way of characterising the goals of health, or of the importance to be placed on such goals (Illich 2002; Fitzpatrick 2001; Habermas 2003).

Likewise we should not downplay the difficulties. The ability to act around a genetic identity is greatly facilitated by (some) knowledge of that identity, and one of the central issues of biomedical big data is precisely the epistemological disempowerment of the data subjects. However, this chapter has argued that it is a general feature of information that it may be understood in different ways in different contexts and by different people – and yet this need not necessarily completely disempower. Even if a completely open situation of 'equal' knowledge and understanding and open reciprocal communication may be an ideal, there

are ways of interaction with less direct communication and alternative ways of acting. There will doubtless be challenges in enabling the data subjects to take part in rational discourse. As data subjects, individuals may be disempowered. As subjects, especially when acting in consort with others, individuals can be immensely resistant and resourceful.

References

Ahmed, T, Lucas, H., Khan, A.S., Islam, R., Bhuiya, A., and Iqbal, M. 2014. eHealth and mHealth initiatives in Bangladesh: A scoping study. *BMC Health Services Research* 14: 260. doi:10.1186/1472-6963-14-260.

Ahmed, T., G. Bloom, M. Iqbal, H. Lucas, S. Rasheed, L. Waldman, A.S. Khan, R. Islam, and A. Bhuiya. 2014. E-*health and M-health in Bangladesh: Opportunities and challenges*. IDS Evidence Report, No 60. Brighton: Institute of Development Studies.

Alpay, Laurence, Paul van der Boog, and Adrie Dumaij. 2011. An empowerment-based approach to developing innovative e-health tools for self-management. *Health Informatics Journal* 7(4): 247–255.

Aristotle. 1999. *Nicomachean ethics*, 2nd edn, trans. T. Irwin. Indianapolis: Hackett Publishing Company Inc.

Armstrong, D., S. Michie, and T. Marteau. 1998. Revealed identity: A study of the process of genetic counselling. *Social Science and Medicine* 47: 1653–1658.

Arribas-Ayllon, Michael, Katie Featherstone, and Paul Atkinson. 2011. The practical ethics of genetic responsibility: Non-disclosure and the autonomy of affect. *Social Theory Health* 9: 3–23.

Atkinson, Paul. 2009. Ethics and ethnography. *Contemporary Social Science* 4(1): 17–30.

Austin, M.A. 2002. Ethical issues in human genome epidemiology: a case study based on the Japanese American Family Study in Seattle. *Washington American Journal Epidemiology.* 1(55): 585–592.

Banks, Sarah. 2015. From research integrity to researcher integrity: issues of conduct, competence and commitment. Academy of Social Sciences. Virtue ethics in the practice and review of social science research. https://acss.org.uk/wp-content/uploads/2015/03/Banks-From-research-integrity-to-researcher-integrity-AcSS-BSA-Virtue-ethics-1st-May-2015.pdf. Accessed 30 June 2015.

Beauchamp, T., and J. Childress. 2009. *Principles of biomedical ethics*, 6th ed. Oxford: Oxford University Press.

Beskow, L.M., W. Burke, J.F. Merz, P.A. Barr, S. Terry, V.B. Penchaszadeh, L.O. Gostin, M. Gwinn, and M.J. Khoury. 2001. Informed consent for population-based research involving genetics. *JAMA* 286(18): 2315–2321.

Blaya, Joaquin, Hamish Fraser, and Brian Holt. 2010. E-health technologies show promise in developing countries. *Health Affairs* 29(2): 244–251. doi:10.1377/hithaff.2009.0894.

Boddington, Paula. 2010. Relative responsibilities: is there an obligation to discuss genomics research participation with family members? *Public Health Genomics* 13(7–8): 504–513.

Boddington, Paula. 2012. *Ethical challenges in genomics research*. Heidelberg: Springer.

Boddington, Paula. 2013. The ethics of emergent knowledge intermediaries. Future Health Systems Innovations for Equality blog. http://www.futurehealthsystems.org/blog/2013/11/13/the-ethics-of-emergent-knowledge-intermediaries.html. Accessed 22 June 2015.

Boddington, Paula, and Maggie Gregory. 2008a. Adolescent carrier testing in practice: The impact of legal rulings and problems with 'Gillick competence'. *Journal Genetic Counselling* 17(6): 509–521.

Boddington Paula, and Maggie Gregory. 2008b. Communicating genetic information in the family: enriching the debate through the notion of integrity. *Medicine Health Care Philosophy* 11(4): 445–454.

Booch, Grady. 2013. The human and ethical aspects of big data. On computing with Grady Booch. IEEE Computer Society. https://www.youtube.com/watch?v=iY7mU1mtQ08. Accessed 24 June 2015.

Busch, L. 2014. Big data, big questions a dozen ways to get lost in translation: Inherent challenges in large scale data sets. *International Journal Communication* 8: 18.

Cassa, C.A., B. Schmidt, I.S. Kohane, and K.D. Mandl. 2008. My sister's keeper? Genomic research and the identifiability of siblings. *BMC Medical Genomics* 1(1): 1.

Clifford, W.K. 1877 [1999]. The ethics of belief. In *The ethics of belief and other essays*, ed. T. Madigan, 70–96. Amherst: Prometheus.

Coiera, Enrico. 2013. Social networks, social media, and social diseases. *British Medical Journal* 346: f3007. doi:10.1136/bmj.f3007.

Crawford, R. 1980. Healthism and the medicalization of everyday life. *International Journal of Health Services* 10: 365–388.

Crawford, Kate. 2013. The hidden biases in big data. *Harvard Business Review*, April 1, 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data

D'Agincourt-Canning, Lori. 2001. Experiences of genetic risk: Disclosure and the gendering of responsibility. *Bioethics* 15: 231–247. doi:10.1111/1467-8519.00234.

Decroo, Tom, Damme Wim Van, Kegels Guy, Remartinez Daniel, and Rasschaert Freya. 2012. Are expert patients an untapped resource for ART provision in Sub-Saharan Africa? *Aids Research Treatment*. doi:10.1155/2012/749718.

Deleuze, Giles. 1988. *Spinoza: Practical philosophy*. San Francisco: City Lights Books.

Dreyfuss, R.C., and D. Helkin. 1992. The jurisprudence of genetics. *Vanderbilt Review* 45(2): 313–348.

Featherstone, Katie, P.A. Paul Atkinson, Aditya Bharadwaj, and Angus Clarke. 2005. *Risky relations*. Berg: Family and kinship and the new genetics.

Fitzpatrick, Michael. 2001. *The tyranny of health: Doctors and the regulation of lifestyle*. London: Routledge.

Floridi, Luciano. 2012. Big data and their epistemological challenge. *Philosophy Technology.* 25(4): 435–437. doi:10.1007/s13347-012-0093-4.

Fricker, Miranda. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Future Health Systems. Future health systems: Innovation for equity. http://www.futurehealthsystems.org/. Accessed 24 June 2015.

General Medical Council. 2009. General medical council guidelines on confidentiality. http://www.gmc-uk.org/guidance/ethical_guidance/confidentiality_contents.asp. Accessed 28 June 2015.

Gitschier, Jane. 2009. Inferential genotyping of Y chromosomes in latter-day Saints founders and comparison to Utah samples in the HapMap project. *The American Journal of Human Genetics* 84: 251–258.

Gregory, Maggie, Rebecca Dimond, Paul Atkinson, Angus Clarke, and Paul Collins. 2007. Communicating about haemophilia within the family: the importance of context and of experience. *Haemophilia* 2007(13): 189–198.

Habermas, Jurgen. 2003. *The future of human nature*. Cambridge: Polity.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, S.F. Nelson, and D.W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics* 4(8): e1000167.

Hume, David. 1955. On miracles. In *An inquiry concerning human understanding*, ed. W. Hendel Charles, 117–141. New York: Bobbs Merrill.

Illich, I. 2002. *Limits to medicine: Medical nemesis the appropriation of health*. London/New York: Marian Boyers.

International HapMap Consortium. 2004. Integrating ethics and science in the International HapMap Project. *Nature Revues Genetics* 5: 467–475.
International Strategy Meeting on Human Genome Sequencing. 1997, 1996. Polices on release of human genome sequence data. http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml
James, William. 1896 [1979]. The will to believe. In *The will to believe and other essays in popular philosophy*, eds. F. Burkhardt et al., 291–341. Cambridge: Harvard.
Keilman, Karina, and Fabian Cataldo. 2010. Tracking the rise of the 'expert patient' in evolving paradigms of HIV care. *AIDS Care* 22(1): 21–28.
Kenen, R.H. 1984. Genetic counselling: The development of a new interdisciplinary occupational field. *Social Science and Medicine* 18(7): 541–549.
Knoppers, Bartha M. 2000. Population genetics and benefit sharing. *Community Genetics*3: 212–214.
Leonard, D.K., G. Bloom, K. Hanson, J. O'Farrell, and N. Spicer. 2013. Institutional solutions to the asymmetric information problem in health and development services for the poor. *World Development Journal* 48: 71–87. doi:10.1016/j.worlddev.2013.04.003.
Lippman, Abby. 1991. Prenatal genetic testing and screening: Constructing needs and reinforcing inequities. *American Journal of Law and Medicine* 17(1–2): 15–50.
Lippman, Abby. 1992. Led (astray) by genetic maps: The cartography of the human genome and health care. *Social Science and Medicine* 35(12): 1469–1476.
Martin, Hammersley, and Paul Atkinson. 1986. *Ethnography: Principles in practice*, 2nd ed. London: Routledge.
Mascalzoni, D., A. Hicks, P. Pramstaller, and M. Wjst. 2008. Informed consent in the genomics era. *PLoS Medicine* 5(9): e192.
Mayer-Schonberger, Victor, and Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*. London: John Murray.
McLean, Bethany, and Peter Elkind. 2004. *The smartest guys in the room: The amazing rise and Scandalous fall of Enron*. London: Penguin.
Mittelstadt, Brent Daniel, and Luciano Florid. 2016. The ethics of Big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–41. doi:10.1007/s11948-015-9652-2.
Montgomery, Jonathan. 2003. Ch. 11. Confidentiality and data protection. In *Health care law*. Oxford: Oxford University Press.
Nagel, Thomas. 1979. Moral luck. In *Mortal questions*, ed. Thomas Nagel, 20–29. Cambridge: Cambridge University Press.
Nissenbaum, Helen. 2004. *Privacy as contextual integrity* (SSRN Scholarly Paper No. ID 534622). Rochester: Social Science Research Network. http://papers.ssrn.com/abstract=534622
Nissenbaum, Helen. 2009. *Privacy in context: Technology, policy, and the integrity of social life*. California: Stanford University Press.
Novas, C., and Nicholas Rose. 2000. Genetic risk and the birth of the somatic individual. *Economy and Society* 29(4): 485–513.
Nutbeam, Don. 2008. The evolving concept of health literacy. *Social Science and Medicine* 67: 2072–2078.
Nyholt, D.R., C.E. Yu, and P.M. Visscher. 2009. On Jim Watson's APOE status: genetic information is hard to hide. *European Journal of Human Genetics* 17(2): 147–149.
Olmen, Josefien, Grace Marie Ku, Raoul Bermejo, Guy Kegels, Katharina Hermann, and Wim Van Damme. 2011. The growing caseload of chronic life-long conditions calls for a move towards full self-management in low-income countries. *Globalization and Health* 7:38. http://www.globalizationandhealth.com/content/7/1/38
Peerson, Anita, and Margo Saunders. 2009. Health literacy revisted: what do we mean and why does it matter? *Health Promotion International* 24: 3. doi:10.1093/healthpro/dap014.
Plows, Alexandra, and Paula Boddington. 2006. Troubles with biocitizenship? *Genomics Society Policy* 2(3): 115–135.
Pope, Alexander. 1796. *An essay on man*. London: Cadell and Davies.
Rabinow, P. 1996. *Essays on the anthropology of reason*. Princeton: Princeton University.

Rachels, James. 1975. Why privacy is important. *Philosophy and Public Affairs* 1: 323–333.
Rose, Nicholas. 2007. *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century*. New Jersey: Princeton University Press.
Rotimi, C., M. Leppart, I. Matsuda, C. Zeng, H. Zhang, C. Adebamowo, I. Ajayi, T. Aniagwu, M. Dixon, Y. Fukushima, D. Macer, P. Marshall, C. Nkwodimmah, A. Peiffer, C. Royal, S. Eiko, H. Zhao, V.O. Wang, J. MCEwan, and HapMap Consortium TI. 2007. Community engagement and informed consent in the International HapMap project. *Community Genetics* 10: 186–198. doi:10.1159/000101761.
Skene, Loanne. 1998. Patients' rights or family responsibilities? Two approaches to genetic testing. *Medical Law Review* 6: 1.
Taussig, K., R. Rayna, and D. Heath. 2003. Flexible eugenics: Technologies of the self in the age of genetics. In *Genetic nature/culture: Anthropology and science beyond the two-culture divide*, ed. A.H. Goodman, D. Heath, and S.M. Lindee. Berkeley: University of California Press.
Toronto International Data Release Authors. 2009. Prepublication data sharing. *Nature* 461(7261): 168–170. http://www.nature.com/nature/journal/v461/n7261/suppinfo/461168a_S1html
Walter, F.M., J. Emery, D. Braithwaite, and T.M. Marteau. 2004. Lay understanding of familial risk of common chronic diseases: a systematic review and synthesis of qualitative research. *Annals of Family Medicine* 2: 583–594.
Wasson, John, Hvitfeldt Forsberg, Staffan Lindblad, Garey Mazowita, Kelly McQuillen, and Eugene C. Nelson. 2012. The medium is the (health) measure: patient engagement using personal technologies. *Journal Ambulatory Care Management.* 35(2): 109–117.
Wellcome Trust. 2003. Sharing data from large-scale biological research projects: a system of tripartite responsibility. http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf
White House, Office of the Press Secretary. 2000. Remarks Made by the President, Prime Minister Tony Blair of England (via satellite), Dr. Francis Collins, Director of the National Human Genome Research Institute, and Dr. Craig Venter, President and Chief Scientific Officer, Celera Genomics Corporation, on the Completion of the First Survey of the Entire Human Genome Project. https://www.genome.gov/10001356. Accessed 6 July 2015.
Williams, Bernard. 1981. Moral luck. In *Moral luck*, ed. Bernard Williams, 20–39. Cambridge: Cambridge University Press.
Wittgenstein, Ludwig. 1967. *Letters to Paul Engelmann*, trans. L Furtmüller. Oxford: Basil Blackwell.
World Health Organisation. 2011. mHealth: New horizons for health through mobile technologies: Second global survey on eHealth. ISBN 978 92 4 156425 0. http://www.who.int/goe/publications/goe_mhealth_web.pdf. Accessed 28 June 2015.
World Health Organisation. 2015. Global observatory for ehealth. http://www.who.int/goe/en/. Accessed 24 June 2015.

Part V
Professionalism and Ethical Duties

Researchers' Duty to Share Pre-publication Data: From the Prima Facie Duty to Practice

Christoph Schickhardt, Nelson Hosley, and Eva C. Winkler

Abstract The purpose of this chapter is to offer an ethical investigation into whether researchers have a duty to share pre-published bio-medical data with the scientific community. The central questions of the chapter are the following: do researchers have a prima facie duty to share pre-published data? And if so, what stakes and aspects of a concrete situation need to be taken into consideration in order to assess whether and to what extent researchers' prima facie duty to share data applies? We will argue that based upon their basic duties to benefit society and to promote scientific knowledge, researchers have a prima facie duty to share data. We will also argue that in order to determine whether the prima facie duty applies in practice it is indispensable to take into account the stakes of the persons concerned as well as context dependent aspects. The chapter's overall goal is to build an analytical and ethical framework that helps to assess with regard to concrete situations whether researchers' duty to share data applies. To this end we analyse the concept of data sharing and clarify what data sharing might imply in practice. To offer an overview of the different stakeholders' concerns we will analyse the normative-informational environment in which data producing researchers (to whom the prima facie duty to share data applies) are usually situated. In the last step we focus on the ethically relevant context dependent aspects and illustrate how they affect researchers' prima facie duty to share data and stakeholders' potentially conflicting stakes.

C. Schickhardt, Ph.D (✉)
University of Heidelberg, Heidelberg, Germany

University of Bamberg, Bamberg, Germany
e-mail: Christoph.Schickhardt@med.uni-heidelberg.de

N. Hosley
Department of Philosophy, Brandeis University, Rabb 303, MS 055, 415 South Street, Waltham, MA 02453, USA
e-mail: nhosley@brandeis.edu

E.C. Winkler, M.D., Ph.D
National Center for Tumor Diseases, University Hospital of Heidelberg, Heidelberg, Germany

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_14

1 Introduction

The purpose of this chapter is to offer an ethical investigation into whether researchers have a duty to share pre-published bio-medical data with the scientific community. The central questions of the chapter are the following: do researchers have a prima facie duty to share data? And, as we will argue that they do, in which conditions and situations does researchers' prima facie duty to share data apply in practice? The overall goal of the chapter is twofold: to analyze and illuminate the conceptual, practical and ethical complexity of a researchers' duty to share data; and to offer an analytical approach and ethical criteria to assess whether and how researchers' prima facie duty to share data applies in concrete situations and in context depending circumstances. In order to reach these goals we proceed in several steps: after introducing recent historical developments and status quo of the ideal and practice of data sharing, we raise the question whether researchers have a prima facie duty to share bio-medical data. We will argue that they have such a prima facie duty. However, we will also argue that in order to determine whether the prima facie duty applies in practice it is indispensable to take into account the stakes of the persons concerned and context dependent aspects of the situation in question. The following sections of the chapter are thus dedicated to propose a conceptual, analytical and ethical basis and framework for determining with regard to concrete situations whether researchers' prima facie duty to share data really applies. To this end we carefully clarify the concept of data sharing and illuminate what data sharing means and implies in practice. We will then explore the different and pertaining claims, rights and duties of stakeholders who are potentially concerned by researchers' duty to share data. In an application oriented last step we carry out an ethical analysis of a typical situation and illustrate the context depending aspects that are ethical relevant and need to be taken into consideration for assessing with regard to concrete situations whether researchers' prima facie duty to share data applies.

Today a main purpose of bio-medical research is to identify and understand causal relations and dynamics of complex, multi-factorial pathological processes and pathways. An increasing number of researchers require massive multi-dimensional data sets to identify relevant causal factors or networks from enormous amounts of non-relevant factors and variants, and to distinguish their pathological meaning for diseases (Bousquet et al. 2014). For instance, to identify genetic drivers in multi-factorial diseases and to draw statistically significant conclusions, researchers need the genomic data and the relevant clinical data of a large sample of patients with similar pathological and clinical background. In order to create such big and multi-dimensional data sets, researchers can aggregate already existing data sets or collect new data and aggregate it with existing data sets. In any case, the sharing of data can be crucial for aggregating data sets in order to reach statistically sufficient amounts of data that are scientifically promising. Data sharing is considered critical for "[r]ealizing the promise of big data" (Kosseim et al. 2014). Despite its potential importance and benefit, data sharing is not a common practice

(Piwowar 2011; Jasny 2013). Given the discrepancy between potential importance and current practice of data sharing, there is a pressing need to explore the question whether researchers have an ethical duty to share pre-published data. To date, this question has enjoyed little attention in the bio-ethical literature. The ethical question of a researchers' duty to share pre-publication data touches one of the most sensitive aspects of every day research. It calls in question the very way in which society and researchers themselves understand researchers' role.

2 Background

Within bio-medical sciences, the field of genomics is widely regarded as the paradigm for data sharing practices (Kaye et al. 2009; Choudhury et al. 2014). It is therefore worthwhile to briefly review data sharing practices in genomics from the Human Genome Project to the most recent developments.[1] The Human Genome Project (1990–2003) was not only formative for genomic sciences, but also with regard to the practices and policies of data sharing. The project established that all sequencing data would be released immediately on the web to be available for the scientific community even before being used in publications, preferably within 24 h (First International Strategy Meeting on Human Genome Sequencing 1996). This policy was described and published in the Bermuda Principles and was further elaborated in the Fort Lauderdale Agreement in 2003 (Fort Lauderdale Agreement. Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility 2003) and the Toronto Agreement in 2009. The policy was also applied by other big projects such as the HapMap Project and the 1000 Genomes project and for certain types of data, by the ICGC (Joly et al. 2012).

Despite the advancements in data sharing, the open access availability of DNA data has been increasingly challenged with publications that have questioned the confidentiality of donors' personal information (McGuire 2008). For instance, Lin et al. (2004) and Homer et al. (2008) cautioned researchers about the identifiabilty of donors through information on single nucleotide polymorphisms. Additionally, in 2013, Yaniv Erlich's team (Gymrek et al. 2013) showed that 50 anonymous male donors who had been sequenced for the 1000 Genomes Project could be identified solely by their DNA sequences and other information freely available on the web. In the aftermath, the risks of re-identification and of non-authorized linkage of genetic and phenotypic or personal data became a major issue of genomic research (Kaye and Hawkins 2014).[2] Currently, restricted access has widely superseded unrestricted open access (Joly et al. 2012; Kaye and Hawkins 2014). Restricted access has become the practice of data archives such as US Data Base of Genotypes and

[1]The following reconstruction is based upon Kaye et al. (Kaye et al. 2009; Kaye and Hawkins 2014).

[2]For a typology of ways of breaching anonymity see Erlich and Narayanan (2014).

Phenotypes (DbGaP), the Wellcome Trust in the United Kingdom and the European Genome-phenome Archive (EGA) (Greenbaum et al. 2011; Shabani et al. 2015). As data access becomes more restricted, data access committees (DACOs) are playing an increasingly important role in evaluating access requests (Joly et al. 2012; Shabani et al. 2015).

Since 2013, the Global Alliance for Genomics and Health (GA4GH) has been promoting global sharing of genomic and health data. GA4GH comprehends many key academic, governmental and private players, including Amazon and Google. The GA4GH initiative's purpose is to develop technical, regulatory and ethical platforms and frameworks in order to promote efficient, safe and responsible global sharing of bio-medical data (Global Alliance for Genomics and Health homepage). In 2003, the US National Institutes of Health (NIH) issued a policy requiring NIH supported Genome Wide Association Studies (GWAS) to release data in the NIH's Data Base of Genotypes and Phenotypes or similar data banks in order to make the data accessible to other users (Simpson et al. 2014). In 2013, NIHs' data sharing policy was extended to Whole Genome and Whole Exome research projects (Simpson et al. 2014). The ideal of making data accessible quickly and in high quality has been adopted in policies by public funders (Kaye et al. 2009; Kaye and Hawkins 2014; Joly et al. 2012). The ideal of data sharing has been implemented in recommendations by research organizations and is increasingly pursued even beyond the sector of genomic research (Poldrack and Gorgolewski 2014; Choudhury et al. 2014) and the bio-medical sciences.[3] An increasing number of funders encourage and require grant holders to share data (Kosseim et al. 2014; Simpson et al. 2014; Kaye and Hawkins 2014; Fortin et al. 2011; Wellcome Trust 2013). Thus, there is growing institutional pressure on data producing researchers to share data and to justify restrictions or exemptions from data sharing. Data sharing has become a de facto condition and contractual obligation for several researchers. However, several barriers and difficulties to data sharing are yet being reported.[4] These include technical, economic, administrational, motivational, ethical and legal barriers as well as barriers concerning governance and oversight by review boards (Dove et al. 2014; Shabani et al. 2015). The ethical exploration of a researchers' duty to share data is pivotal for the question of how funders, society and researchers should cope with these barriers in concrete situations as well as on a more general and institutional level. As we will illustrate in the chapter, on the one hand some barriers might make it challenging, unreasonable, or impossible to ethically require researchers to share data. On the other hand, recognizing a duty to share data would be a reason to motivate and urge researchers to overcome barriers in order to fulfill their duty.

[3]For the United Kingdom see Pryor (2009); Mauthner reports that in the UK making research data available to other users is a central element of the Research Council's remit and of Common Principles on Data Policy issued by Research Councils UK (RCUK) (Mauthner 2013).

[4]For general overviews on barriers to data sharing see van Panhuis et al. (2014), Simpson et al. (2014), Sane and Edelstein (2015).

3 Researchers Prima Facie Duty to Share Pre-published Data

The aim of this section is to explore whether data producing researchers have an ethical prima facie duty to share their data with the research community. First, we need to answer the question of whether there are general ethical principles that justify a researchers' prima facie duty to share data. We think there are two of such principles: the duty to benefit the public and the duty to serve and promote scientific knowledge and scientific progress. As to the ethical justification of these two basic principles, we do not attempt here to provide a (last) ethical foundation. Rather, we will confine ourselves to indicate some reasons why we think it is plausible and reasonable to assume that researchers have a *basic* duty to benefit the public and to serve and promote scientific knowledge and progress.

3.1 Researchers' Duty to Benefit Society and Advance Scientific Knowledge

As to reasons for the duty to benefit the public, researchers' responsibility towards society is comprised of three dimensions: researchers' responsibility as citizens; researchers' responsibility as members of the civil and social elite; and researchers' responsibility as (publicly sponsored) researchers. In terms of principle, the three dimensions of researchers' responsibility towards society should not be viewed as distinct but rather as converging in the person of the researcher as one indivisible moral subject (Nida-Rümelin 2005). Here we will confine our reflections to briefly analysing the question of what researchers as researchers owe to society. One important aspect is that, as we will see below, researchers owe society to serve and promote scientific knowledge and scientific progress. We distinguish researchers' duty to benefit the public from their duty to promote scientific knowledge. Both duties seem to be identical as researchers are mainly supposed to benefit society through scientific discoveries. However, the relation between scientific discoveries and benefit for the public or humanity is not self-evident. For this, it is sufficient to remember the intense discussions beginning in the late 1940s about researchers' responsibilities with respect to the development of nuclear weapons. There can be tensions between the vocation for scientific discoveries and the responsibility for ensuring the common good (Heinemann 2010).[5] The duty to assume responsibilities for society requires researchers to critically reflect on their work's legal, ethical, and political dimensions and implications for society.

To evaluate the effects of research activities on society we can refer to two basic principles: the principle of avoiding harms (and risks), which could be framed

[5]One needs to distinguish the terms "common good" and "public good". We refer to the common good as what benefits society, whereas public good is a specific economical term referring to just one determined and formally defined kind of goods (Bialobrzeski et al. 2012).

as principle of general nonmaleficence, and the principle of increasing societies' welfare, which could be framed as principle of general beneficence.[6] There is growing awareness and support of the requirement that researchers' responsibility is not limited to avoiding harm, but entails the duty to actively benefit society (Meslin and Cho 2010). This holds particularly for publicly sponsored researchers (Heinemann 2010). One obvious reason for this claim is public's ubiquititous and massive financial support of researchers and the research system. "Public funding and support has become an integral part of virtually all research" (Langat et al. 2011). The duty to benefit society applies even more to bio-medical researchers whose explicit social mandate is to benefit the public through scientific contributions to the understanding of health and diseases. We therefore conclude that there are plausible reasons to assume a researchers' basic duty to benefit society. From this assumption, Brakewood and Poldrack (2013) directly conclude that "[t]o provide support for the public trust, researchers have an obligation to share research data with other scientists".

As to researchers basic duty to serve and promote scientific knowledge and scientific progress, one reason to assume such a duty for researchers relies in the ethos of research. By "ethos" we understand "a generally empirically ascertainable, normative structure of role expectations, gratifications and sanctions, guiding convictions, approaches, dispositions and standards that guide the interactions of the respective reference group in which this ethos is effective" (Nida-Rümelin 2005). Scientific knowledge and scientific progress are central values in the ethos of research and constitute a leading ideal for researchers (Nida-Rümelin 2005; Heinemann 2010; DFG 2013, p. 40 ff.). According to their ethos, researchers thus have the basic duty to promote and serve scientific knowledge and scientific progress. Furthermore, society substantially sponsors researchers and research infrastructure. In return, society expects researchers to commit to the general advancement of scientific knowledge. We think that society's expectation towards researchers is well justified. The agreement between society and researchers, and the resulting role of publicly sponsored researchers within society, ground society's right to expect researchers to commit to the advancement of scientific knowledge (Meslin and Cho 2010).[7]

3.2 Data Sharing and the Common Good

Researchers' basic duties to benefit society and to advance scientific knowledge can constitute an ethical basis for researchers' duty to share data – if data sharing

[6]For the principles of nonmaleficence and beneficence (with focus on the relation between physician and patient) see Beauchamp and Childress (2009).

[7]This holds particular for researchers in liberal societies which warrant researchers freedom of research and grant them particular space and authority in the public discourse.

is instrumental to the common good and helps advance scientific knowledge.[8] The second part of this section will be dedicated to some remarks on the relation between data sharing and the common good. Almost all bioethical texts about data sharing display great enthusiasm about the potential advantages and benefits for science and the public. Data sharing is widely considered to advance science and creativity (Kaye et al. 2009; Simpson et al. 2014) and to maximize the utility of data and skills of researchers, accelerating the pace of research and of investigations of specific research questions (Choudhury et al. 2014). According to Kosseim et al. combining and sharing data sets will generate the statistical power needed for accelerated discovery and translate research findings into clinical practice (Kosseim et al. 2014). It might be worth noticing however that in contrast to the bioethical authors, other authors display more awareness with regard to the fact that biomedical "big data" is "big" in relation to currently available technical tools and capacities and that its size and complexity might be challenging or even too big for current computational resources (Mittelstadt and Floridi 2016).

Here we do not mean to fundamentally question the potential of data sharing to benefit the public and to facilitate scientific progress. We will take it for granted that in the field of bio-medical sciences data sharing has a positive and considerable impact on the advancement of scientific knowledge and thus also on the common good. However, in order to justify researchers' duty to share data by reference to the basic duties to benefit society and to advance scientific knowledge, we need to carefully assess the whole picture of potential effects of data sharing. Beyond the intended effects we also need to take into consideration the non-intended side effects, in the short and in the long run. As to the effects on society, one negative and ethically undesirable potential side effect of data sharing might consist in creating additional risks for privacy. Global and large scale data sharing could facilitate the (legal and illegal) collection and analysis of data by internal and foreign governmental intelligence agencies or private industries, and thus contribute to a vital threat to citizens' privacy. Most authors on data sharing mention privacy issues with respect to a single data donor (Sane and Edelstein 2015), but do not address the political and social dimension of privacy. However, privacy is not only valuable for individuals but also for societies for it constitutes an elementary condition for the functioning and flourishing of liberal societies (Rössler 2001). The question whether data can and should be protected against governmental intelligence agencies is hardly ever raised in the literature on data sharing.[9] This seems to be a blind spot in the literature's appeals to the potential of data sharing to benefit the public. Thus, even if one takes it for granted that the scientific advancement of bio-medical

[8] Almost all authors who (briefly) consider data sharing as ethical or scientific imperative, obligation or duty base their claim – among other things – on its (potential) benefit for science and the public (Knoppers et al. 2014; Langat et al. 2011; Brakewood and Poldrack 2013; Chalmers et al. 2014).

[9] For the (legal) importance of the protection against (potential) surveillance by governmental agencies see the safe harbor judgment by the Court of Justice of the European Union (2015).

sciences through data sharing will benefit the public, data sharing has another social-political dimension that relates to privacy and should not be neglected.

To conclude this section, we state that, based upon the basic duties to benefit the public and to promote scientific knowledge and progress, bio-medical researchers have an ethical prima facie duty to share data. This claim relies on the presumption, which seems to be reasonable, that data sharing promotes the common good and the advancement of sciences, even if data sharing might have a (commonly unintentional) negative side effect on privacy that needs to be taken into consideration as well. Thus, the central question of our analysis will be when and under which conditions the prima facie duty to share data applies in practice, in concrete situations. However, we do not propose a definite answer to that question with respect to all potential situations. To evaluate the question whether and how the prima facie duty applies in practice, it is each time necessary to take into consideration the concrete situation (Pearce and Smith 2011). In other words, we believe that the ethically sound answer to the question whether and to what extent researchers' prima facie duty to share data applies is always context dependent and cannot be adequately given in advance for all potential situations. Yet, everybody facing the question whether the prima facie duty applies in a concrete situation will need to have a clear understanding of the concept of data sharing as well as of what data sharing means in practice. That's why the next two sections are dedicated to the clarification of the concept and practical implications of data sharing.

4 The Concept of Data Sharing

The clarification of the exact meaning of the concept 'data sharing' requires addressing several differences. Activities to be considered as 'data sharing' include the transfer or exchange of data between data producers and secondary data users as well as alternative ways of making data available to external researchers (Global Alliance for Genomics and Health 2014). The sharing can occur directly between data producer and data user or indirectly through the intermediation such as data archives or data repositories like, for instance, the NIHs' data base of Genotypes and Phenotypes (dbGaP) or the European Genome-phenome Archive (EGA) (Shabani et al. 2015). A data producer can share his data with one single data user or with several different data users. As mentioned above data can be shared through open access or controlled access. By controlled or managed access the data producer (or intermediaries) can control and select the persons or institutions that will be granted access to the data. Conversely, 'open access' means that everybody has access to the data. Both ways of sharing are compatible with a duty to share data as understood in this chapter. The duty to share as understood in our investigation does not necessarily imply that data be made available to everybody without any control and identification of data users. Yet, it is essential that the data are shared with the whole scientific community. This does not mean that the data is shared

with literally every other scientist. But it requires that every scientist interested in the data can have the data – if he duly submits a data request and meets trust and data protection standards. The duty to share data thus goes far beyond sharing data with trusted and well known colleagues based on mutual trust and common interests (Kaye et al. 2009) or with partners within a research consortium. If a data producer has the duty to share data, he must share it with every researcher and treat every researcher's request for data equally and independently from his own personal interests and preferences. To avoid potential misunderstandings, by a duty to share data we exclude from our account sharing data with data *donors* (patients or research participants). That is, we do not treat questions like the return of health relevant findings or making raw data available to data donors.

As to the term "data", we confine our investigation to data from humans to be used in bio-medical research. We concentrate on bio-medical data from patients and participants of epidemiological studies – who in the following will both be referred to as "data donors". The data that might be shared comprehends different types of data such as genomic data, other omics data and basic data about diagnosis, clinical treatment, and outcome of patients. We also include general personal data about data donors, for instance data concerning the sex and age of data donors, as well as environmental and lifestyle data (even though these data types might be beyond the extension of the term "bio-medical data"). We also need to address the state of scientific elaboration of the data: schematically, there is a scale of elaboration of data in science which goes from raw data at the starting point of scientific analysis to data elaborated in intermediary analysis to final research results. The question of a duty to share data does not pertain to research *results* which are almost always immediately published after being discovered. The question applies to data that still needs to undergo (further) scientific analysis to allow for new scientific insights. So by the duty to share data we refer to raw data, for instance the sequences of Whole Genome Sequencing, and to data resulting from intermediary scientific elaboration (for instance the set of all somatic mutations of a cancer genome, which are the result of the comparison of a cancer patient's germ line genome with the cancer genome but do not constitute research results).

A further crucial point is the distinction between pre-publication data and post-publication data. Our investigation just refers to pre-publication data since we assume that the question whether researchers have the duty to share data that were used for publications, that is post-publication data, is not controversial. It is universally accepted as part of researchers' ethos and codified as good scientific practice that research results and pertaining data are made available to the community for control and reproduction of scientific results (Merton 1961; Campbell et al. 2000, 2002; DFG 2013, p. 43). So the ethical question to focus on is researchers' duty to share *pre*-publication data. The terms "published data" and "non-published data", which are also used sometimes, are slightly misleading since what is usually published is not necessarily all data but just some data. Publications use to contain the research results and do not necessarily include all raw data and all data from intermediate steps which were at the basis of the research results. So the real meaning of the distinction between pre-publication and post-

publication data does not necessarily refer to the comprehensive publication of data but to the question whether the data has been used for making scientific discoveries which were then published. This is worth keeping in mind even if in the field of genomics many journals require authors to make the raw data available through transferring it to a data repository of the journal itself or to a public data archive like the EGA.

Yet, the distinction between pre- and post-publication data is not as easy as one might expect at first glance. According to a study by Campbell et al. "[a]mong geneticists who said they had intentionally withheld information, data or material regarding their published work, [...] 64 % [reported] that they were protecting the ability of a graduate student, post-doctoral fellow or junior faculty member to publish, and 53 % said they were protecting their own ability to publish." (Campbell and Bendavid 2003). The challenge emerging from this citation consists in the fact that researchers might still intend to (partially) use data for (at least one) future publication even after the data was already (partially) used for (at least one) publication. In such a case, notwithstanding a researcher's intention to use data for further future publication, the fact that it has already been used at least once should be considered essential. In the case of an external researcher's request to access data in order to verify (or falsify) the published results, the data should be considered as subject to the widely recognized duty to share post-publication data. However, this does not necessarily imply that secondary data users are entitled to use the data for other goals than the verification of the published results. It is thus necessary to distinguish release of data for verification of published results from genuine data sharing, i.e. release of data to promote secondary data uses.

By the term "data producer" or "data producing researcher" we refer to researchers (or physician scientists) who have data because they have worked to collect it, for example by asking patients for consent to the use of their genetic and clinical data for research activities. By the term "secondary data user" we refer to researchers who ask data producers to make their data available to them. We use the term "secondary data producer" since it seems to be more adequate than the sole term data users. In most cases the data producers use their data for research themselves so that the external researchers who are granted access to the data will not be the first or only ones to use the data, thus making a *secondary* research use of the data.

We also need to differentiate the specific research contexts of both, data producers and secondary data users. We can roughly distinguish a non-profit, academic and/or public research context from a private and for-profit research context. We refer to the first one as "academic" or "public" since it mainly includes universities and publicly sponsored research institutions. However, it also implies non-profit data banks and archives established by patient organizations. The second context will be referred to as "private" since it mainly consists in the industry and companies owned by private stakeholders, for instance the pharmaceutical industry. Since both, data producer and potential secondary data user, may belong to both contexts, data sharing can occur between: (i) academic data producer → academic secondary

data user; (ii) academic data producer → private secondary data user; (iii) private data producer → academic secondary data user; (iv) private data producer → private secondary data user. In this chapter, we will focus on (i), data sharing with both data producers and secondary users working in the academic context. The reasons for this focus are first that a duty to share data appears to be most plausible and feasible in the academic context; second, the participation of private players raises specific ethical and legal challenges concerning for example public-private relationships and property rights of companies. Still, our focus does not mean to suggest that the question concerning the participation of private players in data sharing practices is irrelevant. It is indeed relevant, but also very complicated. For instance, it seems to us that it is much less likely that academic data producer have an ethical duty to share data with private secondary users (case ii) than it is the case with academic secondary users (case i). Furthermore, as a matter of fact (and very concrete interests), it appears to us quite unlikely that private data producers will unconditionally share their data with academic or other private secondary data users – even if the participation of private companies in data sharing initiatives might suggest the contrary.

5 The Practice

After the conceptual clarification of the meaning of data sharing, we now want to shed light on data sharing in practice. We intend to give an idea of what data sharing is like in practice and what single working steps it might imply. We do not claim that all of the steps and elements that we mention are indispensable or essential parts of data sharing. Nor do we claim that a data producing researcher having the duty to share data shall personally carry out all of the steps on his own. Our aim is rather to indicate what data sharing might be like if implemented in an institutional and professional framework and if such implementation is done in a morally conscious way. The following sequential steps of data sharing can be considered as necessary elements of a good practice framework for data sharing.[10]

(i) Consent: Depending on the nature of the data, it is ethically and legally necessary that patients or research participants consent to the (international) sharing of their data. For instance, consent is necessary in the case of sharing genomic data since genomic data cannot be anonymised. If consent is necessary, a 'broad consent' form is most suitable for data sharing and secondary data uses. We will briefly address the problem of broad consent and the ethical debate about broad consent below. Let it here suffice to simply recognize that

[10]Here, we will not mention the very work of collecting and elaborating data which of course is a practical prerequisite for data sharing. We will refer to the work aspect and its ethical relevance in Sects. 6 and 7.

the process of requesting and obtaining broad consent for genomic research including international data sharing is challenging in terms of law, ethics, communication and trust. In many cases before sharing data it must be checked that the data donors' consent cover the planned sharing and secondary uses of their data. If a research project is yet to be designed and data to be collected, it is necessary to adequately address the question of data sharing in the information and consent process.

(ii) Informing the scientific community about the existence of the data: in order to enable scientists (potential secondary data users) from all over the world to require access to data it is a preliminary factual condition that they can learn about the pure existence and availability of the data. Therefore, information about the data must be publicised and available for all scientists in search of data. This holds particularly if the data is to be shared via controlled access and not simply made accessible to everybody on the web via open access. The research community needs to be able to locate information about the existence and nature of the data and about how to submit requirements for data to the data producer.

(iii) Requests for data: if data is shared through controlled access measures, there should be a standardized and efficient procedure for potential secondary data users to submit requests for data (Kaye and Hawkins 2014).

(iv) Deciding about requests for data: there should be an equally standardized and fair procedure for evaluating and deciding about the acceptance or non-acceptance of requests for data sharing. The procedure should address the trustworthiness and data protection standards of the potential secondary data users (Knoppers et al. 2011; Shabani et al. 2015; EURAT 2013). Data producer and secondary data user can submit a Data Transfer Agreement (DTA) which precisely defines the conditions of the agreement.

(v) Technical transfer or access: in the case of controlled access and acceptance of a request for data, the data producer must transfer the data to the secondary user or make it accessible and available to the secondary user in some way.

(vi) Report of breaches and of the end of secondary use: Depending on the Data Transfer Agreement the secondary data user might be obliged to inform the data producer about current uses of data and unforeseen events like confidentiality breaches as well as about the end of the data use and the effective deletion of the data.

Although, as mentioned above, data sharing does not necessarily imply all of these steps, the list of (i–vi) allows us to draw some conclusions. In practice, data sharing is a complicated action that requires technical-computational, administrative and legal steps, and has an important ethical dimension (Simpson et al. 2014). It is unrealistic that all steps implied by data sharing be appropriately and personally carried out by a single researcher or principal investigator. To carefully and responsibly share his data a data producing researcher will need administrative, legal, ethical, technical and infrastructure support.

6 The Normative-Informational Environment of Researchers' Prima Facie Duty to Share Data

So far we have stated researchers' prima facie duty to share data and clarified the concept and practice of data sharing. To assess the question of whether the prima facie duty applies within concrete situations, one must be aware that the data producing researchers (to whom the prima facie duty to share data applies) are already situated within a normative-informational environment.[11] Circumstances in which the question arises whether researchers' prima facie duty applies are not normatively vacuous. There are persons whose claims, rights and duties are potentially concerned by data sharing. Some of these claims, rights and duties can be in tension or conflict with data producers' prima facie duty to share data. The main parties concerned are the data producer himself, the secondary data user, the data donors and the public. In the following, we will refer to these parties as "stakeholders", and will refer to their claims, rights and duties as their "stakes". The purpose of this section is to offer an overview of what the stakeholders' potential stakes to be taken into account in concrete situation might be. By doing so we will also critically discuss claims that might be made by stakeholders, for instance the data producers' potential claim to ownership rights with regard to data. However, here we will not weigh stakeholders' potential claims or other stakes against the prima facie duty to share data.

6.1 The Data Producing Researchers as Stakeholders

Researchers are stakeholders in two different ways: as data producing researchers who have the prima facie duty to share "their" data, and as potential secondary data users. We will start with data producing researchers. Data producers are keen to claim **ownership rights or intellectual property rights** with respect to the data that they collected (Pryor 2009). To justify their property claims they may argue that they have collected the data and that they have invested a considerable amount of work and resources in the collection (and preparation or elaboration) of the data (Langat et al. 2011). Researchers sometimes invest a significant amount of time and resources in the collection and explanation of data. They could appeal to a famous claim by John Locke: that combining one's labor with a natural resource creates natural property rights (Locke 1960, II, chapter V). Thus, in acknowledging the data within the context of the researcher's labor process of data acquisition and explanation, one concurrently acknowledges the researcher's natural rights over the data. On this basis, researchers might conclude that any duty to share pre-publication data is a violation of their property rights and thus unacceptable (Langat et al. 2011).

[11] This approach is inspired by Floridi's concept of infosphere (Floridi 2008).

However, from an ethical perspective there appears to be no sufficient reason and evidence to grant researchers property rights concerning human bio-medical data. First, with respect to personal and sensitive data, data donors still have the right to withdrawal data from research uses even after having consented to participate in data driven research. Second, by consent data donors do not give away property rights but rather authorize persons in their function as researchers to make certain kind of uses of the data for a determined purpose, that is for the sake of research. This is a mandate rather than the transition of property rights. In ethical terms, consent to research by data donors creates stewardship of researchers, not ownership. Third, against the reference to Locke, unlike natural resources in the state of nature, personal data are subject to pertaining ownership and informational rights of the data donors. Fourth, based upon its extensive support to research, society also has claims with regard to research data. We therefore dismiss data producing researchers' claim to create property rights through their work of collecting and elaborating data.

Data producers could abandon ownership claims but insist on some rights, mainly the **rights to have and enjoy the fruits and rewards of their work**. In the life sciences "collecting data can take a number of years" (Pryor 2009). They could argue that the work invested in the collection of the data justifies the following claims: (i) the right to not share the data as long as they have the intention to work on the data in the future; (ii) the right to be granted a defined and limited period of time in which they are allowed to not share the data and which is reasonably long enough to allow them to analyze the data and to publish results (Kaye and Hawkins 2014); (iii) in the case they must share data immediately after collection, data producers might claim the privilege to be the only researchers allowed to publish results of the analysis of the data for a determined period of time which is reasonably long enough to allow them to analyse the data and to publish results. This approach is sometimes called an embargo (Simpson et al. 2014) – even though it is more like a timely limited monopoly; (iv) in any case of sharing data, the data producer could insist on the right of being fairly and adequately rewarded and publicly recognized, for instance through co-authorship, if secondary data users gain and publish results on the basis of the shared data (Kaye et al. 2009; Langat et al. 2011; Kaye and Hawkins 2014); (v) the right to justice and fairness within the institutional system of sciences, including funding and reward: the data producer can claim the right of not being taken advantage of by secondary data users and of being protected against free-riding, as well as the general right to not be disadvantaged within the scientific vocation with its highly competitive character concerning careers, reward, and funding (Langat et al. 2011).

Against these claims (i–v) one might argue that collecting data for research is a job responsibility compensated for by the public. Thus, mere execution of the job cannot ground any particular or additional right beyond the right to a salary. According to this line of thinking, data producing researchers have no rights concerning collected data and can be expected to share data without claiming any further rights to reward, recognition or compensation. If this position was ethically sound, in case of its application, there would still remain practical concerns

about the effects on researchers' motivation to invest time and resources in the collection of data. However, even on an ethical level the position of denying all rights concerning the benefits of their data can be questioned. In particular, the claims to adequate public recognition (v) and to just and fair reward within the institutional system of sciences (v) appear well justified and plausible. One reason backing data producing researchers' claim to be publicly recognized is independent from their personal interests and the logic of careers and funding. According to their ethos and role, researchers are part of the public (Heinemann 2010). They are supposed to publish their results as well as to publish the data used and to publicly give account of their thesis. Public recognition for contributions to scientific discoveries is inherent in both their work and social role. Researchers have a right to be rewarded through public recognition for their contributions to scientific discoveries. It is also a desideratum in the name of accountability and transparency within science as a public discourse. In contrast, the claim to institutional fairness and of not being taken advantage of (v) grounds on (legitimate) personal interests of researchers within the actual system of sciences and the system's internal logic of rewarding and promoting career. Data sharing should neither be prohibitive for career nor for competition for funds such as grants; furthermore, it should not require researchers to sacrifice important resources for their scientific career. As long as the main determinant for career and funding prospects consists in authorship of publications, producing and sharing data must influence career prospects – negatively or positively – only by relation to authorship. However, this might change with the introduction of an additional or alternative standard of attribution and recognition in scientific practice besides an authorship-based standard.

Data producing researchers could also object that a duty to share data violates their right to **freedom of research** (Choudhury et al. 2014). In the German constitution for instance, freedom of research is recognized as a constitutional right of researchers. However, it is questionable that a duty to share data violates the core elements of the freedom of research. Indeed, a duty to share data would not prevent researchers from exercising their autonomy at any stage of research: researchers can proceed freely in choosing their research interest, collecting and analyzing data on a chosen subject, and finally publishing their results.[12]

Data producers can also claim a sort of **second-order-right to not be prevented from fulfilling their duties** towards data donors. As we will illustrate below with regard to data donors as stakeholders, data donors have rights against data producers whom they entrusted with their sensitive data. Data donors have the right that data producers protect their rights to confidentiality and protect them from potential data related risks and harms. Imposing on data producers a categorical duty to share data could put them in ethical dilemmas and urge them to act in a way opposed to what they owe to another stakeholder (Mauthner 2013; Kaye et al. 2009). Data producers thus might sometimes be justified in claiming a second-order-right to an unhindered

[12]But see also Mauthner's critically view of the conflict in the United Kingdom between the Freedom of Information Act and researchers' right to keep data confidential (Mauthner 2013).

path to the fulfillment of their duties (Pearce and Smith 2011). However, there is also the risk that data producers use this claim and the appeal to data protection as false pretense in order not to share data for personal interest.

Last but not least there is the problem of the **costs and burdens of data sharing** (Poldrack and Gorgolewski 2014; Pryor 2009; Sane and Edelstein 2015). Data producing researchers have a fundamental interest in carrying out their own data analysis and research projects without disproportionate hindrance from data sharing. The costs for the human, technical and financial resources used in making data available to secondary data users should be limited and sustainable with regard to their research resources.

6.2 Potential Secondary Data Users as Stakeholders

Potential secondary data users could claim that data producers have the duty to share their data with them. To justify this claim they could refer to the basic duty of all researchers to benefit the public and to promote scientific knowledge. Potential secondary data users might also appeal to a duty of solidarity and reciprocal support among researchers. However, these claims and appeals are unconvincing. Researchers' duty to promote public benefit and scientific progress is justification for granting the potential secondary data users access to data. Yet, it is questionable whether the secondary data users have any *genuine* right to access data. Data producers own the sharing of data (in order to benefit the public and advance science) rather to the public than to secondary data users. The secondary data users thus appear not to have a genuine right towards the data producers. They are rather "instrumental" with respect to benefiting the public and sciences. Similarly, it might be useful to stress that data producers' duty to share data with secondary data users does *not* entitle data users to be the only ones to reap the fruits and rewards of the data. To the contrary since, as researchers, secondary data users too are subject to the researchers' basic duties to benefit the public and to serve and promote scientific knowledge, they have the duty to accommodate the sharing of data and promote the flourishing of trust necessary to data sharing through reliably, honestly and generously sharing the fruits of their data analysis with the data producers. This holds in particular for public recognition, i.e. authorship, for data producers. Last, secondary data users' appeal to reciprocity is objectionable for there is no guarantee of real reciprocity in international data sharing practices among researchers (Sane and Edelstein 2015; van Panhuis et al. 2014).

6.3 Data Donors as Stakeholders

With regard to data donors as stakeholders we first need to remark that the term "data donors" might be somehow misleading since it suggests that all data that

may be shared by data producers has been intentionally "donated" by data *donors*. Contrary to that connotation of the term "data donor", data sharing can also extend to data without patients or research participants having consented to make their data available to data driven research and data sharing. Indeed, the first distinction we need to draw with respect to data donors as stakeholders concerns the ethical and legal requirement of consent for the sharing of determined types of data. Roughly speaking informed consent by persons to the sharing of their data is ethically and legally (depending on the law of each country) necessary as long as the data cannot be efficiently anonymised and as long as the data are sensitive and bear risks of re-identification and potential harms for data donors. As long as data maintain a "personal character" in the sense that they still bear the potential of identification of the data donor, the data donor's rights to informational self-determination and privacy are concerned and his consent is required. By contrast, if data from data donors can be efficiently anonymised, using and sharing the data for research might not concern data donor's rights of privacy and informational self-determination; for this reason it may be ethically and legally legitimate even without data donors' consent. Accordingly in the latter case we assume that data donors have no relevant interests or claims at stake with respect to the sharing of their anonymised data.

We therefore focus on data that requires consent in order to be used for data sharing, for instance genetic data which cannot be anonymised. With regard to the sharing of data that requires data donors' consent, we can schematically distinguish four different groups of data donors: (i) data donors whose data are technically available and could be shared but who were neither informed nor asked for consent to the use and sharing of their data and who thus did not give consent. Listing this sub-group here in the category of data donors who have the right to be asked for consent might appear confusing and counter-intuitive. However, including them here is necessary since there is – at least theoretically – the possibility of using and sharing data from these persons even if this would imply a violation of their rights to privacy and informational self-determination. (ii) Data donors who were adequately informed and asked for consent and gave consent; (iii) data donors who were adequately informed and asked for consent and refused consent; (iv) potential data donors whose data has not been collected yet and who still can be informed and asked to consent to future data driven research and data sharing.

The four subgroups of data donors as distinguished above have the following claims concerning the sharing of their data:

Ad i) On the basis of their rights to privacy and informational self-determination, the data donors who were not asked for consent have the right that their data not be used for research and not shared for secondary data uses. However, we do not want to exclude a priori that there might be highly specific circumstances that could justify the use and sharing of their data even if this would imply the violation of data donors' rights to privacy and informational self-determination.[13]

[13] Yet, at the meta-ethical level we remark that it is in general more challenging to ethically justify a duty which implies the violation of another duty than to justify the legitimacy, that is the allowance,

Ad ii) Data donors who were adequately informed and asked for consent and who gave consent have a couple of rights against the data producer. They have the basic right that their data are used according to what they were told and to what they consented to and not beyond the scope of consent. However, in the case they have given so-called broad consent it might be not always evident if certain kinds of secondary data uses are covered by the consent and by the original intentions expressed by data donors during the information and consent process.

This leads to the much debated problem of the ethical and legal legitimacy of broad consent.[14] In this chapter we do not want to enter into the details of the debate about broad consent since such is not of immediate concern.[15] We confine ourselves to stating that in our view broad consent can be an acceptable tool if handled and interpreted in a certain way and if embedded in an additional normative and governmental framework.[16] Precisely, broad consent for data sharing and secondary data uses should confine the scope of research uses to research activities in the biomedical health sector that comply with good scientific practice and are covered by review boards or similar oversight bodies (Steinsbekk et al. 2013; Henderson 2011). Broad consent should also prescribe measures to protect participant's privacy and warrant the right to withdraw data.[17] More in general, in our view broad consent should be understood as creating a sort of stewardship of data producers. The data

of an act that implies the violation of a duty. In this sense the first question would be whether and under which circumstances a data producer *may* share the data of donors who were not asked for consent. The question whether a data producer has the ethical *duty* to share the data requires particularly strong reasons and justifications.

[14]For an overview on consent with regard to big data driven biomedical research see Mittelstadt (Mittelstadt and Floridi 2015).

[15]Broad consent is in itself a rather broad concept. It might be approximately conceived as lying on a scale between specific consent at the one end and open consent at the other end (Hansson 2009). Whereas traditional consent always addressed and defined the relevant aspects of the study in question, broad consent means consent to a framework of numerous future as well as yet unknown studies (Steinsbekk et al. 2013). Broad consent is a practical and thus attractive way to garner the consent of patients or participants for research that includes secondary uses of their bio-material (tissue) and their data (Hansson 2009). Data sharing to allow secondary data uses is thus not the only topic usually addressed by broad consent and by the discussions about broad consent. However, one might consider data sharing the most delicate point of the broad consent practice. Data can easily be reproduced and transferred and thus bears more risks of abuse and confidentiality breaches for the data donor than the distribution of tissue which is a limited resource. Criticisms of broad consent state that broad consent is not informed consent, that "broad informed consent" is a contradiction in terms (Hofmann 2009), and that the more general consent is, the less informed it is (Árnason 2004). Some conclude that broad consent conflicts with respect for persons and respect for autonomy (Caulfield 2007).

[16]For a similar position see Hansson et al. (2006).

[17]For a theoretical framework of our position concerning governance, researchers' responsibilities and the informed consent process with regard to data sharing, as well for the translation of this position into a code of conduct for non-physician scientists and a consent template form, see EURAT position paper (EURAT 2013).

producers are entrusted by data donors with personal data. Data producers should responsibly assume and exercise the role of stewards with respect to the data donor's intentions.[18] In most cases data donors' main intentions, which potentially come into tension with each other, will be to help science and future patients as well as to have the confidentiality of their data protected. The data producers' stewardship implies the duty to interpret and weigh data donors' intentions in bona fide. The data donors who consented to data driven research including data sharing have therefore the overall right towards the data producer that the latter acts as steward and proxy with respect to their data and related intentions and rights.[19] More precisely, the data producer has the ethical duty to really use the data donors' data for research and to maximize their research use if it was data donors' explicit intention to give a personal contribution to research and to foster scientific progress in the biomedical field. From a legal point of view, consent might be a mere allowance for researchers to use or not to use data for research – without entailing any duty to effectively conduct research. From an ethical point of view, such a formalistic interpretation is unacceptable. Data donors have undergone the entire information and consent process and have expressed their will to promote and assist research by way of entrusting a researcher with the use of their data. Thus, they are ethically justified in expecting the researcher to actively use the data for research and data sharing. From an ethical point of view, broad consent is not only an allowance authorizing a researcher to arbitrarily use or abstain from using the data for research, but also bears a researchers' duty to carry out research and to share the data as described in the consent process. However, data producers do not only owe to data donors to maximize the benefit for research, but also to protect data donors' interests and rights concerning the confidentiality and protection of their personal data.

Ad iii) Persons who were informed about data sharing and asked for consent and who explicitly denied consent have an inviolable right that their refusal be accepted and respected and that their data remain unshared. Respect for one's autonomous decision to refuse consent is at the very core of the right to autonomy, the right to be respected as persons and, more precisely, of the rights to privacy and informational self-determination.

Ad iv) Potential data donors have, of course, the right to be adequately informed and asked for consent. Depending on whether they give or refuse consent, they are identical with the subgroups (ii) and (iii).

[18] Brakewood and Poldrack speak of a "fiduciary relationship" between researcher and a research participant (Brakewood and Poldrack 2013).

[19] It might be philosophically inspiring to link the bioethical concept of stewardship for biomedical data to a more general concept from Floridi's information ethics, that is to his concept of creative stewards as referring to moral agents within the infospshere (Floridi 2008).

6.4 The Public as Stakeholder

The main stake of the public with respect to data sharing has been mentioned above when discussing whether data producing researchers have a prima facie duty to share data. We stated there that publicly sponsored researchers have the basic duties to benefit the public and to promote scientific knowledge and scientific progress. The public's rights have a crucial importance in the ethical justification of these basic duties which, in turn, justify data producers' prima facie duty to share data. The public has the right toward the data producer that the data producer share data in order to benefit the public and to advance scientific knowledge. We can remark that this ethical framework is also the reason why it is the public and not secondary data users who has the right toward the data producer that the data producer share data with secondary data users.

Like the data producers themselves, the public too could claim **ownership rights or intellectual property rights** with respect to the data collected by "its" publicly sponsored researchers. The public can argue that data producers are public employees and that any result of their work including data sets is property of the public. According to Guttmacher et al. "[t]he model of the investigator owning data has been increasingly replaced by one in which society owns data." (Guttmacher et al. 2009). Based upon the public's substantial investments in research and data generation, for Langat et al. "research data is more accurately described as common property" (Langat et al. 2011) of researchers and the public. We reject a comprehensive ownership claim by the public or state for reasons we referred to above when rejecting ownership claims by data producers. Furthermore, viewing data producers as mere representatives of the state would imply that the state has the duties and responsibilities with respect to data donors that we attributed to the data producers themselves: the overall duty to assume stewardship and act as bona fide proxy with regard to data donors' data, intentions and pertaining rights. We conclude therefore that pre-publication data are neither property of the public or the researchers, nor their common property.

A last remark on claims of the public as stakeholder concerns the relations between different "publics", that is, between societies of different countries (Langat et al. 2011). A public that extensively sponsors collection of data through researchers might have the claim of not being taken advantage of by the public of another country who does not sponsor data collection or whose researchers do not equally share data. Poor countries have a particular right to benefit from the use of data that was collected by its researchers or stems from its people. A just distribution of the fruits of scientific results enabled by data sharing might be put at risk by differences in the scientific and political capacities of states and their national research systems.

To conclude, this section's purpose was to offer an overview about the stakeholders concerns, that is their claims, rights and duties potentially concerned by data producers' prima facie duty to share data. In each situation where the question arises whether data producers' prima facie duty applies, the stakes and the prima

facie duty have to be weighed and balanced against each other. To do this, it is also necessary to take into account some aspects which are ethically relevant for weighing the single stakes and the prima facie duty and which vary from situation to situation. Among these aspects the most important one is the potential benefit from sharing concrete data sets for the public and for the advancement of scientific knowledge. In the next section we will give an overview about these aspects. We will exemplify how, in terms of principle, they affect the ethical weight of the single claims, rights, and duties at stake and thereby impact the evaluation whether the prima facie duty applies.

7 Towards Practice: Exemplifying the Analytical and Evaluative Approach

In each concrete situation, the question whether researchers' prima facie duty applies is rather complex. The complexity is due to the following factors: (i) there are different stakeholders; (ii) the different stakeholders have different claims, rights, and duties of concern to researchers' prima facie duty to share data; (iii) the stakeholders' stakes are partially in tension or conflict among each other and, most importantly, with regard to the data producers' prima facie duty to share data; (iv) there are general aspects (circumstances) such as the potential benefit of data sharing, which are ethically relevant and vary from situation to situation; (v) the ethically most important of these aspects, such as the potential benefit or the risks of data breaches and harms for data donors, are difficult to assess and predict; (vi) the ethically relevant aspects can almost infinitely vary and differ from situation to situation.

Due to this complexity we argue that the question whether data producers' prima facie duty applies in practice can only be adequately evaluated and answered by studying precisely the concrete situation in question. In this section we focus on a specific type of situation that we believe to be quite common but which however is only a schematically constructed type of situation and not a concrete situation. By treating this kind of situation we first intend to exemplify how to analyze and identify stakeholders' relevant stakes; second, we propose a list of the different context dependent aspects which are ethically relevant and requiring consideration; we third aim at shedding light on what these aspects in terms of principle mean for the ethical evaluation of the stakeholders' stakes and the question whether the prima facie duty to share data really applies.

More precisely, we will focus on the kind of situation in which the data producer has data from data donors, in which consent for data sharing is necessary because the data is sensitive and cannot be fully anonymised, and in which data donors have been adequately informed and asked for consent and have given consent to participate in data driven research including the sharing of their data for secondary uses. As to the stakeholders claims, rights and duties in our analysis we will only include

those which are most relevant and ethically well justified according to our above mentioned analysis of the stakeholders' potential claims. For instance, we will not mention data producers' ownership claims concerning the data, for we have argued above that this claim is ethically unjustified.

In the defined kind of situation the data producer can make the following claims concerning the enjoyment of the fruits of the research results that were enabled by data sharing:

(i) to not share the data so long as they have the intention to work on the data in the future;
(ii) the right to be granted a defined and limited period of time in which he is allowed to prevent data sharing, and which is reasonably long enough to allow him to analyze the data and to publish results;
(iii) in case he must share data immediately after collection, to be granted the privilege to be the only researcher allowed to publish results of the analysis of the data for a determined period of time;
(iv) in any case of sharing data, to be warranted the right of being fairly and adequately rewarded and publicly recognized;
(v) to be granted justice and fairness within the institutional system of sciences, especially regarding institutional funding and reward;
(vi) as to the costs and burdens of data sharing, to not be obliged to spend a disproportionate part of her time and resources in making available data to other researchers.

As to the claims and right of the other stakeholders, the data producer has the following duties:

(i) most importantly, he owes the public the prima facie duty to share his data;
(ii) towards data donors he has the duty to act as a steward of their data, will and interests. This implies the duty towards data donors to maximize the research benefit of their data through sharing it, and at the same time to adequately protect the confidentiality of their data.

A closer look at the data producer's claims reveals potential tensions and plain conflicts with his prima facie duty to share data. The claim not to share data as long as he is still willing to work on it in the future (i) or not to share data for a limited period of time (ii) are in clear conflict with his prima facie duty to share pre-publication data as owned to the public and the data donors. Data producers' claims (i) and (ii) are in fact tantamount to an almost unconditioned (i) or a conditioned claim (ii) not to share data. Data producers' claim to be granted a monopoly (embargo) with respect to publishing results from the data for a determined period of time after sharing data (iii) means to share data but to condition the secondary uses of it by prohibiting secondary users to publish their results before the end of the data producers' publication monopoly. Secondary data users can work on the data during the embargo time and publish their results the day after the end of the embargo. For the data producers this means to carry out the

analysis of the data under a considerable time pressure (Kaye et al. 2009) and that even if they succeed in publishing results from the data, publications by secondary data users presenting results of the same data might follow within short time (and thus reduce the value of the data producers chapter in the scientific literature). The further claims by data producers are not in plain conflict with the prima facie duty to share data. This holds for the claims to public and fair recognition (iv), to not be taken advantage of and not be disadvantaged in the institutional system of sciences (v), and to not be disproportionately burdened or hindered from carrying out the analysis of his data by the costs and burdens from making the data available to secondary data users (vi). However, these claims can easily come into tension with the prima facie duty to share data if they are not adequately satisfied. The more the duty to share data implies that the data producer foregoes what he deserves the more the application and justification of the duty to share data are ethical challenging. If the data producers' claims (iv–vi) are not adequately and reasonably satisfied, the prima facie duty to share data is likely not to apply or to have limited binding force (for example as supererogatory duty).

To assess the ethical weight of stakeholders' stakes and to weigh them against data producers' prima facie duty to share data, it is necessary to take into consideration the following ethically relevant aspects of the situation in question:

(a) Benefit: Benefit is the most important criterion since it is central to researchers' basic duties which justify their prima facie duty to share data. In our context the criterion of benefit refers to the *potential* benefit from the analysis of data and the scientific results. The equivalents to the potential benefit of data are the opportunity costs for the public benefit and the advancement of scientific knowledge that are caused by not sharing the data. The potential benefit (or the opportunity costs) of data sharing are difficult to predict or to measure. In most cases, the data producers themselves or the experts of the scientific community, including the potential secondary data users, are likely to have an idea about the scientific value and potential of data. Objective criteria like the amount and statistical power of a data collection, the quality and the level of elaboration of the data or the rarity of the data in the research community might be helpful for assessing the potential benefit of data. The amount of time during which data is likely not to be shared, or within which no secondary data user is allowed to publish results of his analysis of the data, might serve as one objective, even though not sufficient, criterion for assessing the opportunity costs of keeping data private. The ethical direction in which the benefit criterion points is the following: The greater the benefit (or the corresponding opportunity) costs, the stronger data producers' duty to share data (as owned to the public and to the data donors). Put in other terms, the greater the potential benefit, the more the prima facie duty does apply to data producers in concrete situations.

(b) Work invested in the collection of the data: it is plausible that the amount of work invested by data producers to collect (and elaborate) data has impact on

the justification of their claims (i–v). The direction in which the criterion of work points is the following: the greater the amount of work and efforts invested by data producers, the more justified their claims concerning their rights to the fruits and rewards of the scientific results earned by way of the data. We believe that this holds particularly for their right to fair recognition (iv) and to systematic fairness (v).

(c) Costs and burdens for data producers: the costs and burdens associated with all working steps necessary to share data with secondary users should be limited, reasonable and sustainable with respect to the data producers' research resources. The direction in which this criterion points is the following: the greater the costs and burdens for the single data producer's research resources, the more it is ethically challenging to justify requirements for data sharing.

(d) The risks for data donors: The sharing of data donors' data with secondary data users on a potentially global scale implies risks for data donors with respect to the protection and confidentiality of their data and of potential harms from the abuse of their data. Even though it is difficult to measure and predict the long term risks implied by data sharing, there are several criteria to refer to when assessing the data risks for data donors: the trust worthiness of the secondary data user and of his institution, the quality of accountability and data protection standards of secondary data user's institution, and the data protection law to which the secondary data user and his institution are subject. The general direction in which this criterion points is: the greater the presumed risks for data donors, the more weight the data producer has to attribute to data donors' interest in protection of their data and the less weight to data donors' intention to promote science. Also, based upon data producers' second-order right not to be prevented from keeping to their duties and responsibilities, the greater the risks for data donors, the less the public has a right to require the data producer to share data.

Further circumstances to be taken into consideration when evaluating the application of the prima facie duty within concrete situations are the followings:

(e) the costs for the public caused by the data producers' collection of data: the greater the costs from data sharing sustained by the public, the stronger the justification of public's claim that the data producer should share data to benefit the public.

(f) the potential therapeutic benefit from data sharing for the individual data donor; the potential therapeutic benefit can add to the data producer's duty towards the data donor to share his data;

(g) the availability of alternatives to the ethically problematic sharing of data from data donors: before making ethically problematic use of data, on should examine whether there are feasible and less problematic alternatives.

8 Conclusion

This chapter is meant to serve as systematic and foundational ethical exploration into the question whether researchers have a duty to share pre-publication bio-medical data with the scientific community. Data sharing is widely considered crucial and promising for the advancement and acceleration of bio-medical knowledge and the scientific understanding of health and diseases. However to date, data sharing is not commonly practiced and institutionalized. This raises the question whether researchers have an ethical duty to share data. We argue that since researchers have the basic duties to be of benefit to society and serve and promote scientific knowledge, they have a prima facie duty to share data. However, on a social and political level bioethicists and data sharing initiatives should also take into consideration the potential (and unintended) threat that might result from data sharing to citizen privacy. The ethical evaluation of the question whether and when researchers' prima facie duty to share data applies in practice is a highly complex matter. The complexity is mainly due to two circumstances: first, there are several parties that are stakeholders and whose claims, rights and duties might be in conflict with researchers' prima facie duty to share data; second, there are ethically relevant and situation dependent aspects. We emphasize that to give an ethically sound and balanced evaluation of the question whether in a concrete situation a researcher has the duty to share his data, one needs to confront the complexity with a robust ethic, one that can account for the stakeholders' concerns and other additional features of context. To help everybody to evaluate whether and to what extent researchers' prima facie duty to share data applies within a particular situation, in this chapter we offer conceptual clarifications, an analytical overview of stakeholders potential claims, rights, duties, as well as a list of general and ethically relevant features of context which need to be taken into account.

From our exploration, we draw also some rather pragmatic conclusions. First, with regard to authorities or funding institutions ready to impose a duty to share data on researchers, we stress that, due to the normative complexity and data producers' right not be prevented from keeping to their responsibilities, researchers' should be granted the possibility to have their say. This should enable them not only to make claims on their supposed rights, but also to draw attention to their responsibilities and duties towards data donors which might be (partially) incompatible with data sharing. Second, there are several aspects which may come into tension with researchers' prima facie duty to share data but could and should be systematically addressed by the state, scientific institutions and the scientific community. If these aspects were adequately addressed, researchers' prima facie duty to share data would apply more easily and within more concrete situations. The state could introduce laws that protect data donors against attempts of re-identification and against the abuse of their data. The state and funding institutions could increase funding for data sharing infrastructures and for remediating costs fromdata sharing.

The state, institutions and the scientific community could implement measures and governance frameworks in order to ensure that in case of data sharing data producers' justified claims are really, certainly and adequately respected.

Acknowledgement Funding: The first author's contribution to the article was supported by a funding (01GP1404A) of the German Federal Ministry of Education and Research. Acknowledgments: We thank Sebastian Schleidgen for commenting on an earlier draft, Prof. Dr. Stefan Fröhling and Prof. Dr. Rudi Balling for inspiring discussions, and Simone Dippel for support in literature research.

References

Árnason, Vilhjálmur. 2004. Coding and consent: moral challenges of the database project in Iceland. *Bioethics* 18(1): 27–49. doi:10.1111/j.1467-8519.2004.00377.x.

Beauchamp, T.L., and J.F. Childress. 2009. *Principles of biomedical ethics*. Oxford/New York: Oxford Univ. Press.

Bialobrzeski, A., J. Ried, and P. Dabrock. 2012. Differentiating and evaluating common good and public good: Making implicit assumptions explicit in the contexts of consent and duty to participate. *Public Health Genomics* 15(5): 285–292.

Bousquet, J., C. Jorgensen, M. Dauzat, A. Cesario, T. Camuzat, et al. 2014. Systems medicine approaches for the definition of complex phenotypes in chronic diseases and ageing. From concept to implementation and policies. *Current Pharmaceutical Design* 20(38): 5928–5944.

Brakewood, Beth, and Russell A. Poldrack. 2013. The ethics of secondary data analysis: Considering the application of Belmont principles to the sharing of neuroimaging data. *NeuroImage* 82: 671–676. doi:http://dx.doi.org/10.1016/j.neuroimage.2013.02.040.

Campbell, Eric G., and Eran Bendavid. 2003. Data-sharing and data-withholding in the genetics and the life sciences: Results of a national survey of technology transfer officers. *Journal of Health Care Law and Policy* 6(2): 241–255.

Campbell, Eric G., Joel S. Weissman, Nancyanne Causino, and David Blumenthal. 2000. Data withholding in academic medicine: Characteristics of faculty denied access to research results and biomaterials. *Research Policy* 29(2): 303–312.

Campbell, E.G., B.R. Clarridge, M. Gokhale, et al. 2002. Data withholding in academic genetics: Evidence from a national survey. *JAMA* 287(4): 473–480. doi:10.1001/jama.287.4.473.

Caulfield, Timothy. 2007. Biobanks and blanket consent: The proper place of the public good and public perception rationales. *King's Law Journal* 18(2): 209–226. doi:10.1080/09615768.2007.11427674.

Chalmers, Donald R.C., Dianne Nicol, and Margaret F. Otlowski. 2014. To share or not to share is the question. *Applied and Translational Genomics* 3(4): 116–119. doi:http://dx.doi.org/10.1016/j.atg.2014.09.011.

Choudhury, Suparna, Jennifer R. Fishman, Michelle L. McGowan, and Eric T. Juengst. 2014. Big data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience* 8: 239. doi:10.3389/fnhum.2014.00239.

Court of Justice of the European Union: Press Release 117/15. 2015. Judgement in Case C-362/14 Maximillian Schrems v Data Protection Commissioner, October 6. http://curia.europa.eu/jcms/upload/docs/application/pdf/2015-10/cp150117en.pdf. Accessed 18 Nov 2015.

DFG. 2013. *Sicherung guter wissenschaftlicher Praxis. Denkschrift. Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft"*, 2nd ed. Weinheim: WILEY-VCH.

Dove, Edward S., Bartha M. Knoppers, and Ma'n H. Zawati. 2014. Towards an ethics safe harbor for global biomedical research. *Journal of Law and the Biosciences* 1(1): 3–51. doi:10.1093/jlb/lst002.

Erlich, Yaniv, and Arvind Narayanan. 2014. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* 15(6): 409–421.

EURAT (Ethical and Legal Aspects of Whole Human Genome Sequencing). 2013. Position Paper. Cornerstones for an ethically and legally informed practice of Whole Genome Sequencing: Code of Conduct and Patient Consent Models. http://www.uni-heidelberg.de/md/totalsequenzierung/informationen/mk_eurat_position_paper.pdf. Accessed 07 Nov 2015.

First International Strategy Meeting on Human Genome Sequencing. 1996. Bermuda principles. http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1. Accessed 02 Nov 2015.

Floridi, Luciano. 2008. Foundations of information ethics. In *The handbook of information and computer ethics*, ed. Kenneth E. Himma and Herman T. Tavani, 3–23. Hoboken: Wiley.

Fort Lauderdale Agreement. 2003. Sharing data from large-scale biological research projects: A system of tripartite responsibility. http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf. Accessed 04 Nov 2015.

Fortin, S., S. Pathmasiri, R. Grintuch, and M. Deschênes. 2011. 'Access arrangements' for biobanks: A fine line between facilitating and hindering collaboration. *Public Health Genomics* 14(2): 104–114.

Global Alliance for Genomics and Health. 2014. Framework for responsible sharing of genomic and health-related data. https://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data. Accessed 02 Nov 2015.

Greenbaum, Dov, Andrea Sboner, Mu Xinmeng Jasmine, and Mark Gerstein. 2011. Genomics and privacy: Implications of the new reality of closed data for the field. *PLoS Computational Biology* 7(12), e1002278. doi:10.1371/journal.pcbi.1002278.

Guttmacher, Alan E., Elizabeth G. Nabel, and Francis S. Collins. 2009. Why data-sharing policies matter. *Proceedings of the National Academy of Sciences of the United States of America* 106(40): 16894. doi:10.1073/pnas.0910378106.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339(6117): 321–324. doi:10.1126/science.1229566.

Hansson, M.G. 2009. Ethics and biobanks. *British Journal of Cancer* 100(1): 8–12. doi:10.1038/sj.bjc.6604795.

Hansson, Mats G., Joakim Dillner, Claus R. Bartram, Joyce A. Carlson, and Gert Helgesson. 2006. Should donors be allowed to give broad consent to future biobank research? *The Lancet Oncology* 7(3): 266–269. doi:http://dx.doi.org/10.1016/S1470-2045(06)70618-0.

Heinemann, Thomas. 2010. Forschung und Gesellschaft. In *Forschungsethik. Eine Einführung*, ed. Michael Fuchs, Thomas Heinemann, Bert Heinrichs, Dietmar Hübner, Jens Kipper, Kathrin Rottländer, Thomas Runkel, Tade Matthias Spranger, Verena Vermeulen, and Moritz Völker-Albert, 98–119. Stuttgart/Weimar: J.B. Metzler.

Henderson, Gail E. 2011. Is informed consent broken? *American Journal of the Medical Sciences* 342(4): 267–272.

Hofmann, B. 2009. Broadening consent—and diluting ethics? *Journal of Medical Ethics* 35(2): 125–129. doi:10.1136/jme.2008.024851.

Homer, Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics* 4(8), e1000167. doi:10.1371/journal.pgen.1000167.

Jasny, Barbara R. 2013. Realities of data sharing using the genome wars as case study – An historical perspective and commentary. *EPJ Data Science* 2(1): 1–15. doi:10.1140/epjds13.

Joly, Yann, Edward S. Dove, Bartha M. Knoppers, Martin Bobrow, and Don Chalmers. 2012. Data sharing in the post-genomic world: The experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Computational Biology* 8(7): e1002549. doi:10.1371/journal.pcbi.1002549.

Kaye, Jane, and Naomi Hawkins. 2014. Data sharing policy design for consortia: Challenges for sustainability. *Genome Medicine* 6(1): 4. doi:10.1186/gm523.

Kaye, Jane, Catherine Heeney, Naomi Hawkins, Jantina de Vries, and Paula Boddington. 2009. Data sharing in genomics – Re-shaping scientific practice. *Nature Reviews Genetics* 10(5): 331–335. doi:10.1038/nrg2573.

Knoppers, Bartha Maria, Jennifer R. Harris, Anne Marie Tassé, Isabelle Budin-Ljøsne, Jane Kaye, Mylène Deschênes, and Ma'n H. Zawati. 2011. Towards a data sharing Code of Conduct for international genomic research. *Genome Medicine* 3(7): 46. doi:10.1186/gm262.

Knoppers, Bartha M., Jennifer R. Harris, Isabelle Budin-Ljøsne, and Edward S. Dove. 2014. A human rights approach to an international code of conduct for genomic and clinical data sharing. *Human Genetics* 133(7): 895–903. doi:10.1007/s00439-014-1432-6.

Kosseim, Patricia, Edward S. Dove, Carman Baggaley, Eric M. Meslin, Fred H. Cate, Jane Kaye, Jennifer R. Harris, and Bartha M. Knoppers. 2014. Building a data sharing model for global genomic research. *Genome Biology* 15(8): 430. doi:10.1186/s13059-014-0430-2.

Langat, Pinky, Dmitri Pisartchik, Diego Silva, Carrie Bernard, Kolby Olsen, Maxwell Smith, Sachin Sahni, and Ross Upshur. 2011. Is there a duty to share? Ethics of sharing research data in the context of public health emergencies. *Public Health Ethics*. doi:10.1093/phe/phr005.

Lin, Z., A.B. Owen, and R.B. Altman. 2004. Genetics. Genomic research and human subject privacy. *Science* 305(5681): 183.

Locke, John. 1960. *Two treatises of government*. Cambridge: Univ. Press.

Mauthner, Natasha. 2013. Open access data sharing policies: Implications for academic roles, practices and identities. Society for Research into Higher Education.

McGuire, Amy L. 2008. Identifiability of DNA data: The need for consistent federal policy. *The American Journal of Bioethics* 8(10): 75–76. doi:10.1080/15265160802478511.

Merton, Robert K. 1961. *Social theory and social structure*. Glencoe/Illinois: The Free Press.

Meslin, Eric M., and Mildred K. Cho. 2010. Research ethics in the era of personalized medicine: Updating science's contract with society. *Public Health Genomics* 13(6): 378–384. doi:10.1159/000319473.

Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.

Nida-Rümelin, Julian. 2005. Wissenschaftsethik. In *Julian Nida-Rümelin*, ed. Angewandte Ethik, 835–860. Stuttgart: Alfred Kröner.

Pearce, Neil, and Allan H. Smith. 2011. Data sharing: Not as simple as it seems. *Environmental Health* 10: 107. doi:10.1186/1476-069X-10-107.

Piwowar, Heather A. 2011. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE* 6(7), e18657. doi:10.1371/journal.pone.0018657.

Poldrack, Russell A., and Krzysztof J. Gorgolewski. 2014. Making big data open: Data sharing in neuroimaging. *Nature Neuroscience* 17(11): 1510–1517. doi:10.1038/nn.3818.

Pryor, Graham. 2009. Multi-scale data sharing in the life sciences: Some lessons for policy makers. *International Journal of Digital Curation* 4(3): 71–82.

Rössler, Beate. 2001. *Der Wert des Privaten*. Originalausg., 1. Aufl. Aufl. Suhrkamp-Taschenbuch Wissenschaft; 1530, vol. 1530. Frankfurt am Main: Suhrkamp.

Sane, Jussi, and Michael Edelstein. 2015. *Overcoming barriers to data sharing in public health. A global perspective*, ed. Centre on Global Health Security. Chatham House.

Shabani, Mahsa, Knoppers Bartha Maria, and Borry Pascal. 2015. From the principles of genomic data sharing to the practices of data access committees. *EMBO Molecular Medicine* 7(5): 507–509. doi: 10.15252/emmm.201405002.

Simpson, Claire L., Aaron J. Goldenberg, Rob Culverhouse, Denise Daley, Robert P. Igo, Gail P. Jarvik, Diptasri M. Mandal, et al. 2014. Practical barriers and ethical challenges in genetic data sharing. *International Journal of Environmental Research and Public Health* 11(8): 8383–8398. doi:10.3390/ijerph110808383.

Steinsbekk, Kristin Solum, Lars Øystein Ursin, John-Arne Skolbekken, and Berge Solberg. 2013. We're not in it for the money—Lay people's moral intuitions on commercial use of 'their' biobank. *Medicine, Health Care and Philosophy* 16(2): 151–162. doi:10.1007/s11019-011-9353-9.

van Panhuis, Willem G., Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J. Herbst, David Heymann, and Donald S. Burke. 2014. A systematic review of barriers to data sharing in public health. *BMC Public Health* 14: 1144. doi:10.1186/1471-2458-14-1144.

Wellcome Trust. 2013. Impact of the draft European Data Protection Regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtvm054713.pdf. Accessed 28 Nov 2015.

Reporting and Transparency in Big Data: The Nexus of Ethics and Methodology

Stuart G. Nicholls, Sinéad M. Langan, and Eric I. Benchimol

Abstract Examples of biomedical big data are routinely-collected health data. These may include information collected in electronic health records (EHRS), disease registries, or health administrative datasets. The ability to use this information for research has raised important questions regarding security and confidentiality. However, we suggest that a neglected area of discussion pertains to post-analytic aspects of research using biomedical big data. Specifically, there has been a lack of attention paid to the ethical obligation of transparent and complete reporting of studies using large-scale health-related datasets.

In this chapter we argue that improving the transparency and quality of reporting is ethically important for a number of practical as well as principled reasons. From a practical perspective the accurate reporting of methods allows for appropriate peer review and critical evaluation of studies; facilitates reproduction and replication of research findings; may help to reduce waste, and avoid redundancy and unnecessary repetition; and may facilitate public trust in scientific research. We may also have

S.G. Nicholls (✉)
Children's Hospital of Eastern Ontario (CHEO) Research Institute, Research Building #1, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada

School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
e-mail: snicholl@uottawa.ca

S.M. Langan
London School of Hygiene and Tropical Medicine, London, UK
e-mail: Sinead.Langan@lshtm.ac.uk

E.I. Benchimol
Children's Hospital of Eastern Ontario (CHEO) Research Institute, Research Building #1, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada

School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

Department of Pediatrics, University of Ottawa, Ottawa, ON, Canada

Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada
e-mail: EBenchimol@cheo.on.ca

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_15

principled reasons to improve the reporting of studies using routinely-collected health data; reasons that may relate to researcher integrity and population benefits such as the fair use of resources, minimising risk of harms, and maximising benefits.

We conclude this chapter by presenting the recommendations from the RECORD (REporting of studies Conducted using Observational Routinely-collected Data) Statement (http://record-statement.org/), an international collaboration involving stakeholders using routinely-collected health data, together with a reflection on the way in which reporting guidelines improve the quality of reporting and where there is still more to do.

1 Introduction

Our capacity to conduct ever-more data intensive research has increased dramatically in recent decades. Advances in data capture technology and in computing power has allowed researchers to conduct analyses that were impractical in the recent past. In the health sphere, biomedical big data includes data as diverse as imaging information, phenotypic data, molecular or other 'omic data as well as clinical, environmental, and behavioural information that may be routinely-collected, but which may also result from specific projects (Wang and Krishnan 2014). These data are usually characterised by the four V's: *volume*, that is the amount of data produced; *velocity*, that is the speed of data collection; *variety*, the different types of data sources; and *veracity*, which relates to the quality of the data – how credible is it? (Council of Canadian Academies 2015; Mittelstadt and Floridi 2016; Nuffield Council on Bioethics 2015).

An example of such biomedical big data are routinely-collected health data (Council of Canadian Academies 2015). These may include information collected in electronic health records (EHR), disease registries, or health administrative datasets. Data such as records of outpatient health care contacts, hospitalisations, or surgeries are collected for administrative and clinical purposes, without specific *a priori* research questions (Dean et al. 2009; Harpe 2009). However, there is a growing interest in using these data in a research context. The draw of using these datasets include: the opportunity to identify effects that individually would be small, but may be substantial at the population level and would hitherto have been difficult to identify in small, single study samples (Currie 2013); that such datasets can assist with the identification of specific subgroups at particular risk and with sufficient power, when previously such analyses would have proved impossible; and they provide a population-based picture, especially of interest in jurisdictions with universal health care.

Given these potential draws, routinely-collected health data are being used for observational studies, comparative effectiveness research (Hoffman and Podgurski 2013), health services research, and clinical trials (Dean et al. 2009; Maeng et al. 2014; Kreuter et al. 2014), as well as in research to develop decision-rules that would improve clinical efficiency (predictive analytics) (Cohen et al. 2014). Indeed,

a number of funding bodies, such as the Canadian Institutes of Health Research, have endorsed the use of administrative health databases for outcomes research and in particular as a strategy for enhancing patient-oriented research and improving health care efficiency and delivery (Canadian Institutes of Health Research 2011).

The ability to link large scale information has, rightly, raised important questions regarding security of information, as well as patient privacy and confidentiality (Mittelstadt and Floridi 2016; Lane et al. 2014). However, we suggest that a neglected area of discussion pertains to post-analytic aspects of research using biomedical big data. Specifically, there has been a lack of attention paid to the need for transparent and complete reporting of studies using large-scale health-related datasets. Considering the imperfect nature of routinely-collected health data, and of the analytic tools used, there is significant risk that consumers of research may be unaware of hidden sources of bias. In order to avoid the trap of assuming that such data are a 'black box' and to adhere to the scientific tenants of reproducibility and replicability, adequate reporting must be ensured.

Accurate and complete reporting is necessary in order to evaluate the clinical utility of findings (Cohen et al. 2014; Kreuter et al. 2014), but also brings forth a number of ethical issues (Vayena et al. 2015). In the remainder of this chapter we argue that improving the transparency and quality of reporting is a key issue for studies using biomedical big data, such as linked health data, and outline important ethical drivers.

The rest of the chapter proceeds as follows. After a brief orientation to the topic of transparency and quality of reporting, we outline key arguments for the transparent reporting of studies making use of biomedical big data. We then outline the underlying ethical values and principles that support the need for accurate and complete reporting of studies, before discussing the development of the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement.

2 Transparent Reporting in Health Research: Practical Implications

Several studies indicate that the reporting of health-related research is inadequate. Chalmers and Glasziou, for example, have suggested that with respect to research evidence relevant to clinicians and patients, over 50 % of studies are never published in full (Chalmers and Glasziou 2009). Further, they suggest that 40 % of reports of clinical trials do not contain adequate information on interventions, although this may be reduced to 10 % by checking references, contacting authors, and doing additional searches (Chalmers and Glasziou 2009). Poor reporting, they contend, can make the information unusable and add up to billions of dollars being wasted (Glasziou et al. 2014). We suggest that the motivation for improving the reporting of research using routinely-collected health data can be grouped into the following areas: appropriate peer review and identification of bias such that they may impact on the perceived utility and implementation of findings; the advancement of

scientific understanding through replication and furthering of research that builds upon previous discoveries (Lazer et al. 2014a); maximization of limited health-research resources by decreasing study redundancy brought about by unnecessary repetition, and; facilitation of trust in research through openness to scrutiny (Vayena et al. 2015). Indeed, these issues are potentially magnified with biomedical big data: the sheer scale of the datasets will mean that there may be limited opportunities for replication and analogous data collection in other contexts (Kreuter et al. 2014).

2.1 Peer Review and Evaluation

Peer review and the evaluation of publications for methodological soundness is a cornerstone of the scientific process. This step not only relates to pre-publication peer review but also ongoing post-publication critical evaluation of research. The ability to accurately evaluate data quality, methods, and results within a manuscript is fundamental to this process, and may have important consequences: if authors do not provide sufficient detail concerning the conduct of their study, then reviewers – and subsequently other readers – will have only a partial understanding of the process undertaken. As a consequence it would be difficult to evaluate the methodological quality or judge the reliability of the results, interfering with the interpretation of findings and their appropriateness (Moher et al. 2009, 2010; Collins and Tabak 2014). The issues with interpretation might include evaluation of potential biases (O'Connor 2010), lack of applicability to context, or simply poor methodological rigor. These potential issues may be exacerbated in the context of data that isn't collected with a research intent (Harpe 2009). Indeed, with routinely-collected health data there are a number of aspects that are important to evaluate, and which directly pertain to the fourth V of the big data definitional quartet: veracity.

Questions of veracity may be multiple: there may be data errors (Guttmann et al. 2010), missing data (Kreuter et al. 2014), or data that are out of date (Weiskopf and Weng 2013). Moreover, as Hoffman notes, the volume of information in large-scale biomedical datasets and potential linkage between datasets introduce "myriad of opportunities for the introduction of errors and omissions" (Hoffman and Podgurski 2013). These errors and omissions are magnified when linkage methods are not accurate or complete (Cohen et al. 2014).

As an example of this complexity, consider a simple research study that seeks to identify a cohort of individuals with a specific condition, and look at their health outcomes and health services use. Given the question, health administrative data would be a useful resource, but it may be unclear how well the condition in question was documented, even with existing coding schemes such as the International Classification of Diseases (ICD). Researchers may need to develop a diagnostic algorithm that can be used to accurately identify individuals with a pre-specified condition (Benchimol et al. 2011), as opposed to relying on the innate coding within the dataset. Consequently, in order to evaluate how accurate the analysis of the data is with respect to the prevalence or incidence of the disease, one needs to know

how accurate the identification of individuals within the dataset was. Failure to report the algorithm and results of validation work, or to provide citation of the validation procedure, would mean that readers are unable to ascertain the risk of misclassification bias (Manuel et al. 2010), and thus whether the final results should be trusted and used in decision-making.

The veracity may also be affected by selection bias in the information contained by a resource. For example, Jacobs et al., report on the publication of physician-level data from Medicare on the Consumers' Checkbook website (Jacobs et al. 2009). These data provide physician ratings for surgical procedures based on the data in the Medicare database. In the example provided by the authors, an individual searching this dataset may be misled as to the applicability of the ratings to their own case: the Medicare databases is largely restricted to patients over 65 and likely represents only a fraction of physician encounters or procedures. Thus physicians who conduct more procedures in other populations may be underrepresented in these limited datasets, and individuals' consulting reports based on these data may make inappropriate or ill-informed decisions if these limitations of the dataset are not made clear.

Missing data may also occur, even though the variables are present in the dataset. For example, blood pressure may be an important risk factor for coronary heart disease (Franklin et al. 2001). If patient data on blood pressure is missing – even though blood pressure is a field for which data should be collected – then analyses that exclude the variable or that have large amounts of missing data in the field, may provide erroneous results. Moreover, if the missing data is associated with some other characteristic of the individual or population, then there may be a selection bias that could provide misleading information. For example, if men are for some reason less likely to have their blood pressure recorded, then erroneous conclusions may be drawn about the role of blood pressure in health risks when compared to women. There is also the possibility of unmeasured confounding, in which a variable is not present in the dataset (and therefore cannot be included in multivariable analyses). This residual confounding could alter the results of the analysis of association between predictor and outcome.

Thus, with respect to methodological evaluation of studies, reporting transparency is important, as it will allow critical evaluation of the study and its strengths and biases. When a lack of transparent reporting prevents the appropriate and necessary review of research, then readers will be required to make assumptions which may result in incorrect conclusions.

2.2 Reporting, Reproduction, and Replication

Another core aspect of health research is the ability to reproduce or replicate research findings. Transparent reporting not only facilitates decision making by all readers, but allows for the reproduction or replication of methods by interested researchers (Collins and Tabak 2014). Adequate documentation of process is important for both and is a core standard of reporting (Wager and Kleinert 2011).

Reproduction refers to the ability of other researchers to conduct the same analyses and generate the same research findings on the same dataset (Thompson and Burnett 2012). Being able to create the same results from the same data is important from the perspective of consistency – inconsistent results may point to variation in the methods or data that could call into question the validity of the results obtained. Several studies point to difficulties in reproducing research findings (Vasilevsky et al. 2013), with a number citing poor reporting of methods for the inability to reproduce results. This, it is argued also "prevents us from retrospectively tagging a resource as problematic or insufficient, should the research process reveal issues with a particular resource" (Vasilevsky et al. 2013). That is, the poor reporting of methods and subsequent inability to reproduce findings prevents researchers from identifying if the source of variation lies in the methods or the data source.

Replication, in contrast, allows other researchers to reconstruct a study using new data or in a new cohort in order to evaluate the robustness of findings. The ability to replicate a study may corroborate a proposed causal explanation, or may substantiate a scientific claim. Equally, the inability to replicate the findings of a study may point to errors or shortcomings in the original study, or to the use of different methodologies or algorithms (O'Connor 2010; Cook and Collins 2015). Replication is also important for evaluating the generalizability of research findings: can the same methods be applied to different data sources and yield the same findings? How applicable are the present results in other populations?

Transparent reporting is similarly important when study findings cannot be replicated. A failure to replicate a study may point to some additional and important element that is missing. This may be important theoretical aspects, or contextual knowledge, that is required in order to replicate previous research findings. This relates to what Feynman refers to as 'cargo cult science' (Feynman 1998) – that in addition to the availability of procedural steps there also needs to be an understanding of the context in which the research was undertaken. Accurate and complete reporting of studies may, therefore, not solely be about transparency regarding the process of research, but also understanding the context in which the content exists. Thus, in the context of routinely-collected health data, it may be important to know how data are collected and the purposes for which they were collected, in order to understand the potential for replication. Complete and accurate reporting of research is therefore a precursor to replication.

An illustrative example where a lack of transparency has stymied replication (and indeed methodological evaluation), is that of Google Flu Trends (GFT). GFT was launched in 2008 and used data mining techniques to assess trends in flu-related searches on Google.com in order to predict outbreaks of influenza (Butler 2013). While studies indicated initial results from GFT were promising (Cook et al. 2011), surveillance data collected by the US Centers for Disease Control (CDC) has indicated that in recent years, and despite revisions in Google's algorithms, the predictions have been highly inflated in comparison with actual cases (Butler 2013; Lazer et al. 2014a, b). Further raising concerns is the opacity of reporting by the developers of GFT. This is, apparently, partly motivated by concern that divulging

the list of search terms used in the algorithm would adversely affect the model. Developers and researchers have noted that when terms have been published by Google as part of their Google Blog, the frequency of searches for these terms have increased dramatically (Copeland et al. 2013). The difficulty with such an approach by Google, as discussed by Lazer and colleagues, is that "it would be impossible to replicate the analyses of the original paper from the information provided regarding the analysis."(Lazer et al. 2014a)

2.3 Reducing Waste, Avoiding Redundancy and Unnecessary Repetition

While the ability to replicate and/or reproduce research is a desirable outcome of transparent reporting, and is increasingly supported by scientific publishers seeking to support scholarly research (Glasziou et al. 2014), unnecessary repetition may be an undesirable outcome of poor reporting, potentially leading to wasted research resources (Ioannidis 2012).

Poorly reported studies are also associated with an opportunity cost. For every study that is inadequately reported, the opportunity is lost to fund an alternative study whose results may have been reported more optimally, and whose results may yield greater health benefits than those that were funded but were subsequently poorly reported.

With routinely-collected health-data, transparent and accurate reporting may have additional benefits that reduce waste. For many sources of routinely-collected health data there is no research oversight; that is, data custodians may require approval processes, but there is not associated review of study questions to ensure efficient use of the data and prevent unnecessary duplication of analyses (Friedman 2007; Ioannidis 2012). Transparent reporting may help to reduce duplication by clearly detailing the datasets and methods used, and which may identify sources of redundancy in proposed analyses, and thus reduce waste in resources.

2.4 Transparency of Reporting and Trust in Scientific Research

Transparent and complete reporting of studies that have made use of routinely-collected health data provide a method for establishing trust between users of research publications and individuals or organisations who use the data. By publishing studies in a way that clearly shows what data were used, in which contexts and ways, and by whom this was done, there is a greater openness regarding how health data are collected, used, and the outcomes of the research conducted using routinely-collected health data – information that the public has been shown to want (Sciencewise 2014; Damschroder et al. 2007). Indeed, some research points to

public concerns not being so much about the collection of data, but rather the ends to which it is put (Sciencewise 2014). This relates not only to the use of aggregated and anonymised data compared to individual level data, but also the types of analyses being undertaken. For example, research has pointed to concern by some groups over the entities with which data is shared, including the sharing of data between government departments if it is felt that those analyses may be used in ways that would be the detriment of some social groups.

Indeed, a failure in transparency can potentially derail research using routinely-collected health data. An example of this is the recent debate surrounding the UK use of routinely-collected health data as part of *care.data*. Under *care.data* individual patient information from National Health Service (NHS) records across different NHS organisations could be brought together for research purposes (Nuffield Council on Bioethics 2015), potentially generating results that could lead to improved healthcare delivery. Concern was raised about the potential for health records to be used in ways that might reveal personal information, and tension developed over the use of records by profit-seeking entities. Responses to these concerns appeared not to be satisfactory, in part due to an apparent lack of transparency regarding who would have access to these data and the purposes for which research would be conducted (Taylor 2014).

As such, the opaque nature of the process by which data were to be collected and used served to undermine the trust in the organisation acting as custodians of the data, and the lack of information regarding the uses to which the data would be put further eroded this trust. Hence, while appropriate and transparent reporting may serve the purpose of allowing individuals to assess the quality of research using biomedical big data, a corollary is that in doing so the research process is opened up. Opening up the research process may facilitate public awareness of both the data being used for research and the purposes for which it is being used, but also the results of the research – which may engender trust in researchers and research institutions.

3 Principles, Values and the Transparent Reporting of Research

While the above examples indicate practical reasons why the accurate and complete reporting of studies is important, the ethical grounding of the need for transparent reporting of health research is largely left implicit.[1] Several declarations and codes

[1]Our focus here is limited to steps to improve the transparency and completeness of reporting as part of an effort to improve standards of reporting. As such, we do not engage in discussion of deliberate acts of research misconduct, such as the falsification or fabrication of research results. In the subsequent sections we limit our discussion to the former, but acknowledge that biomedical big data also potentially presents important issues pertaining to the latter – particularly with respect to the ease with which data may be manipulated – that require further consideration.

of practice make explicit statements that reporting should be complete and accurate (Last 1996), but the reasons are not articulated. For example, the Declaration of Helsinki states that:

> 36. Researchers, authors, sponsors, editors and publishers all have ethical obligations with regard to the publication and dissemination of the results of research. Researchers have a duty to make publicly available the results of their research on human subjects and *are accountable for the completeness and accuracy of their reports. All parties should adhere to accepted guidelines for ethical reporting* (emphasis added) (World Medical Association 2013).

Further, and citing the ICH Harmonised tripartite guidelines, Needleman and colleagues argue that the appropriate reporting of results should be considered as a requirement for transparency (Needleman et al. 2008).

These guidelines of course raise two questions: what are the principles guiding this ethical obligation, and against what standards should researchers and publishers be held? Despite the above statements of the obligations of researchers, and increasing interest in the need for transparent reporting of research (Altman and Moher 2014; Anderson et al. 2013; Groves 2008; Moher 2007, 2008; Simera et al. 2010a, b; Beauchamp 1996), surprisingly few authors or institutions have articulated in detail what these obligations are, and from where they derive. In the next section we address the former, offering examples of how the need for increased transparency and completeness of reporting of research using biomedical big data derives from important ethical values, before reporting on the development of the RECORD Statement (record-statement.org), an international project to develop reporting standards for routinely-collected health data.

3.1 Individuals and Research Integrity

When one thinks of the values or principles that have motivated the desire to improve the reporting of studies making use of routinely-collected health data (and biomedical big data more generally), one can point to two domains: those values or principles that derive from values intrinsic to individuals as those who conduct research, and those that derive from external sources and which relate to the ends to which research results are put. While these motivations may not be specific to the question of completeness of reporting for routinely-collected health data, the sheer complexity of big data in health and potential limited opportunities for replication in other settings, mean that there is a greater emphasis placed on the requirements for improved reporting of studies.

As part of this former domain in which the focus is on values intrinsic to individuals, Davenport et al., under the auspices of the expert group on research integrity for the Canadian Council of Academies, locate appropriate reporting within a broad research integrity framework. They argue that there are five core values that drive approaches to improve research integrity: honesty, fairness, trust,

accountability, and openness (Davenport et al. 2010). This conception of locating reporting transparency within a broader research integrity framework is consistent with what Anderson and colleagues have referred to as the 'research ethics lifecycle' (Anderson et al. 2011). This perspective focuses on values pertaining to individual integrity: that is, the values reflect assessments of the individuals conducting and reporting the research, such that these become assessments of individual character. Thus, the onus is placed on the individual to be honest in their methodological reporting evidenced by complete and transparent reporting and to which they are held accountable.

Honesty A component of this is honesty of the individual. Most obviously this would relate to personal obligations not to deliberately deceive readers, but this may also refer to what Masic terms 'intellectual honesty'; of being straightforward with the description of the research hypothesis, analysis, interpretation, and reporting of results (Masic 2012).

Fairness With respect to fairness, researchers benefit from the work of others (Institute of Medicine 2015) from which they draw ideas, methods, or data and on which they can build. As such, a principle of fairness and reciprocity requires them to also contribute to this pool of knowledge so that others may benefit (Carter et al. 2015). Fairness also relates to the content of the publication and the obligation of researchers not to be biased in their selection of evidence or presentation of findings. Such a process has the practical benefits – as indicated earlier – of facilitating replication or avoiding unnecessary repetition of research thus moving forward the scientific knowledge base in an efficient manner.

Trustworthiness, accountability, and openness Being transparent in process and practice, characterised by visibility or accessibility of information, is possible through the publication of research (Davenport et al. 2010). Complete and accurate publication of research may facilitate trustworthiness in, and accountability of, individuals under both of what Onora O'Neill refers to as the *audit agenda* and the *openness agenda* (O'Neill 2002). Under the former, openness may improve accountability by facilitating monitoring, inspection, or audit of performance – in effect one is able to audit publications through peer review and evaluation. If discrepancies or deficiencies are identified, then these may be put to the authors and they may be held accountable.

Under the openness agenda, the availability of information is not so much used to hold individuals or institutions to account, but as a marker which indicates the trustworthiness of said individual or institution and it is the potential for scrutiny that then acts as a motivator. Thus, by proactively reporting studies using routinely-collected health data in a way that is consistent with best practice, an individual or institution provides an indicator of their potential trustworthiness. Acting in concert, the accurate and complete publication of research opens the researcher to scrutiny (as part of an audit agenda), such that one may gain evidence of their competence (or lack thereof), which then inspires confidence that the researcher/institution

is trustworthy (Manson 2010). Moreover, the ability and willingness to write in detail about a study may also be taken as indicating that the researcher (or research institution) is honest.

3.2 Extrinsic Values and Implications for Research

Others have focused on values with respect to impact external to the individual researcher. In large part this focus derives from the fact that interest in biomedical big data is not actually about the data itself, but rather the ends to which the data – and more specifically results from analyses of the data – are put (King 2016). In their assessment of the ethical challenges of big data in public health, Vayena and colleagues take a broader perspective and cite as important values: risk of harm, fair use of resources, and trust, transparency, and accountability (Vayena et al. 2015).

Risk of harm If the reporting of research conducted using routinely-collected health data is poor, then important findings may be overlooked due to uncertainty regarding the methods used to obtain them. As a result, resources may be allocated to sub-optimal care or in developing healthcare policy that would provide sub-optimal outcomes. Perhaps more disconcertingly, results may be taken up into practice or policy when doing so is unwarranted and this may expose patients to unnecessary risk of harm (Needleman et al. 2008).

Nonmaleficence is often invoked in the context of clinical research and used in relation to risk of harm to study participants. In the broader sense used by Vayena et al. (2015), the concern pertains to the potential for harm brought about by inappropriate uptake of research findings. To this we suggest that the concept of beneficence should be added: we should also seek to maximise the uptake of findings that will improve health outcomes, and the reporting of research should be allied to this. Indeed, these are commonly cited arguments within the research community who have pushed for increased transparency of reporting for medical research generally (Institute of Medicine 2015). Replication – and by implication the reporting of research that facilitates replication – may increase the likelihood that beneficial results of research will be taken up into practice or healthcare policy as the evidence base increases (Institute of Medicine 2015).

Justice A failure to transparently report research, and thus impede reproduction or replication, is important from a perspective of distributive justice, particularly if the failure to replicate research leads to unwarranted burden on some members of society. For example, groups may be inappropriately identified as being affected by disease – and consequently stigmatised – when a failure to report transparently impedes replication studies that could potentially illustrate limitations in the original study (Vayena et al. 2015). Being able to verify the veracity of the data will be essential in such circumstances – was there a systematic bias in the original data? Was there uncontrolled confounding? – in order to ascertain whether the findings

reflect an accurate representation or deficiencies within the data. Transparent and complete reporting of not only the study results, but also the methods and data sources, will be imperative for such an assessment. This could lead to healthcare resources being used inefficiently – either through unnecessary stockpiling of vaccine, or conducting additional research and surveillance at additional costs for both society and the individuals under surveillance. Alternatively, potentially useful information that could be generated – either directly or through scientific advances that build on the unreplicated studies – will be forgone and harm may be incurred that could otherwise be prevented. Thus, the need for verification of data quality and clear articulation on potential sources of bias (as previously discussed) have clear implications for concerns over the ends to which research results are put in the context of routinely-collected health data.

Fair use of resources Justice is also relevant to the fair use of resources and broader community who may be impacted by the conduct or results of a study. There are opportunity costs to funding research – even if studies using routinely-collected health data are less costly than studies with primary data collection. Consequently, funding a study that duplicates an existing, but poorly reported study means that other research is not funded. Such unfunded research could yield potentially useful results and as such, there may be benefits that are foregone by not undertaking this research. Poorly reported studies may also precipitate a waste of resources as other researchers attempt to build on findings, or unnecessarily repeat studies, when otherwise they would not have done so if they were able to appropriately critique the methods or assumptions (Vasilevsky et al. 2013). Consequently, principles of fairness to society in the way that resources are used require the complete and accurate reporting of studies such that resources are not wasted.

Trust, transparency and accountability While trustworthiness, transparency and accountability are consistent in principle with individual-level attributes, they also exist in relation to a broader societal perspective. As Elliot et al., have argued:

> [S]ociety is likely to be better served when scientists strive to be as transparent as possible [. . .] Transparency can promote public trust by helping laypeople understand how both empirical evidence and value assumptions enter into scientific decision making and policy formation (Elliott and Resnik 2014).

This is highly pertinent in big data research using routinely-collected health data where individual consent to participate is not usually feasible (Beauchamp 1996), and where the open nature of data collection means that specific information on the purposes of data collection and analysis cannot be established in anything more than a broad sense (ter Meulen et al. 2011).

Improved reporting of research may allow researchers to fulfill what Carter and colleagues have called the "social licence" of research. This social licence, they argue, represents a need to go beyond mere compliance with legal mandates but also to act in a way that is consistent with the expectations of society regarding the conduct of those activities, with such a licence granted to certain occupations (such as academic or health researchers) (Carter et al. 2015). Obtaining, and maintaining,

trust not only legitimises the use of resources to conduct research, but also facilitates the uptake of beneficial research findings to improve health.

In summary, there may be values intrinsic to the researchers themselves – those that are embodied by ethical individuals – but also values that are associated with the end use of research, such as the use of resources and trust in scientific research. While these ethical drivers have been established, and statements to the effect that publication in accordance with accepted guidelines have been made, routinely-collected health data has thus far been lacking in accepted standards. In the final section of this chapter we outline the development of such a reporting standard: the RECORD statement (Benchimol et al. 2015; Nicholls et al. 2015).

4 Applying Ethics: Reporting Guidelines as an Approach to Improving the Reporting of Research

Despite the ethical drivers, the quality of reporting of research – on an international scale – using routinely-collected health data has been suboptimal (de Lusignan and van Weel 2006; Hemkens et al. 2015; Benchimol et al. 2011), a finding consistent with a general trend within the biomedical research literature (Kilkenny et al. 2010).

An approach to improving the reporting of research is the development of reporting guidelines. Reporting guidelines have generally been developed for specific study types; there is the CONSORT guideline for the reporting of Randomised Controlled Trials, STROBE for the reporting of observational studies (von Elm et al. 2007), STARD for diagnostic studies (Bossuyt et al. 2003), PRISMA for systematic reviews and meta-analyses (Moher et al. 2009), as well as a host of other examples (Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network). These generally take the form of a checklist of items that should be included in publications of a particular type. These checklists often include explanatory text designed to assist authors in reporting the specific type of research. Notably, they tend to focus on the necessary elements of the study methods and results that require explicit discussion (Altman and Simera 2014).

Reporting guidelines are usually developed through a consensus method approach in which experts in the relevant area are brought together, often after the conduct of a systematic review of the literature, to develop a 'best practice' guideline. As such the development of guidelines most closely aligns to the aforementioned intrinsic aspects of "Trustworthiness, accountability, and openness", given the development by a body of researchers who would be reporting studies. As such they take on a form of self-regulation, in which those researchers working as experts in a particular field develop a set of *de facto* standards or professional norms.

Checklists are often endorsed by journals, either through pronouncements of support in editorials or instructions to authors, or through requirements that authors submit a completed checklist to indicate that they have complied with an appropriate

reporting guideline (Altman and Moher 2014). Some journals may also publish a copy of a completed checklist alongside an article. The argument here being that doing so increases transparency (Altman and Simera 2014). As such, while the guideline development may be partly a response to those intrinsic drivers, their use as documentary evidence of compliance is in line with O'Neill's notion of audit and the external validation of the research. In doing so, there is clear application of the external principles of "trust, transparency and accountability", itself in keeping with the broader desire for accessibility to enhance peer review. Indeed, a noted potential benefit of compliance with reporting guidelines has been to facilitate the conduct of systematic reviews and meta analyses through the better documentation of elements that may introduce bias or confounding (Altman and Moher 2013; Altman and Simera 2014).

Several studies have now explored the impact of reporting guidelines on the transparency and completeness of study reporting. These studies suggest that citation by a particular study of a reporting guideline improves adherence to the checklist in use, but also that endorsement by a journal also leads to improved adherence to reporting of items within studies published within the journal (Turner et al. 2012; Sorensen et al. 2013; Armstrong et al. 2008; Moher et al. 1999; Prady et al. 2008).

Research conducted using health administrative data and other routinely-collected health data typically falls under the broad rubric of 'observational studies' from a methodological standpoint: they are non-randomised. Observational studies, as opposed to intervention studies, have been subject to the development of their own reporting guideline: the **STR**engthening the **R**eporting of **OB**servational studies in **E**pidemiology **(STROBE)** statement (von Elm et al. 2007). An impetus for STROBE was research noting that epidemiological studies often failed to report basic methodological components such as the eligibility criteria for the study, or how cases and controls were identified. The aim, therefore, was to develop a checklist of items that should be reported in cohort, case-control, and cross-sectional studies; the main types of observation study within epidemiology (von Elm et al. 2007).

However, a number of issues generated by the use of routinely-collected health data are not addressed by STROBE. For example, an important aspect for interpretation of research conducted using routinely-collected health data would be a description of the database characteristics (the population, inclusion and exclusion criteria), as well as diagnostic codes and algorithms to identify exposures and outcomes (Langan et al. 2013). Given these deficiencies, there are significant difficulties for researchers wishing to comply with pronouncements that they publish in accordance with accepted guidelines; the poor fit of STROBE to routinely-collected health data means that there are no accepted standards thus creating a difficulty for researchers using this type of data in complying with their ethical mandate.

5 Setting Standards for Reporting: Improving Our RECORD

Given the limitations of existing guidelines a guideline development process was undertaken to expand the existing STROBE guidelines with additional criteria relevant to the reporting of studies using routinely-collected health data. In accordance with guidance on the development of reporting guidelines (Moher et al. 2010), a three-stage development process was used. First, two modified electronic Delphi surveys of key stakeholders were undertaken (Nicholls et al. 2015). Survey participants included, but were not limited to, clinicians, clinical and academic researchers, journal editors, policymakers, and pharmaceutical industry representatives. The first survey was qualitative in nature and sought to identify themes that were deemed important to include in the RECORD statement. The second survey took the themes derived from the first survey and asked participants to prioritise them for inclusion within the final RECORD statement. The Delphi was followed by a second stage consisting of a face-to-face meeting of the RECORD working committee members. During the meeting, committee members reviewed the survey results together with free-text comments that were also provided. Following review and discussion, a checklist of items and explanatory text were created. The third and final stage of the process involved further open comment on the draft checklist, conducted through an online message board on the RECORD website (record-statement.org).

The final results are presented in Table 1. Recommendations from the stakeholders, and finalised by the working group reflect three broad areas of concern and perceived need for improvement: the identification of research using routinely-collected health data; the evaluation of important methodological components; and information regarding the accessibility and limits imposed on the data used. By way of contrast, there was little comment regarding the reporting of results suggesting that much of the necessary requirements of studies using routinely-collected health data are covered by existing guidance in STROBE. Again, these findings suggest that the main need for transparency and completeness of reporting relates to how the research was done, which relates back to the practical and principled elements of being able to evaluate and replicate research.

These elements closely map onto the above practical and principled reasons for accurate and transparent reporting of studies using routinely-collected health data.

Notably, an important aspect of the guidelines related to the ability to identify a study as one using routinely-collected health data. Considering the increasing interest in research using routinely-collected health data it is noteworthy that at present there is a lack of Medical Subject Heading (MeSH) terms with which to search for these types of studies within the main electronic databases. As such, studies using routinely-collected health data are not being consistently catalogued such that a range of keyword search terms may be used in order to find relevant studies. This not only makes it difficult for individual researchers to find relevant studies, it also poses a problem for those conducting systematic reviews. Systematic reviews – which, along with meta analyses serve as filtered research sources – sit at the top of the evidence pyramid, a hierarchy of types of evidence that forms the

Table 1 The RECORD statement checklist of items

	Item number	STROBE items	RECORD items
Title and abstract			
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.
		(b) Provide in the abstract an informative and balanced summary of what was done	RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract.
			RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated
Introduction			
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	
Objectives	3	State specific objectives, including any pre-specified hypotheses	
Methods			
Study design	4	Present key elements of study design early in the paper	
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	
Participants	6	(a) *Cohort study* – Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.
		Case–control study – Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and Controls.	RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.

		Cross-sectional study – Give the eligibility criteria, and the sources and methods of selection of participants	RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.
		(b) *Cohort study* – For matched studies, give matching criteria and number of exposed and unexposed	
		Case–control study – For matched studies, give matching criteria and the number of controls per case	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.
Data sources/measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	
Bias	9	Describe any efforts to address potential sources of bias	
Study size	10	Explain how the study size was arrived at	
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	
Statistical analyses	12	(a) Describe all statistical methods, including those used to control for confounding	
		(b) Describe any methods used to examine subgroups and interactions	
		(c) Explain how missing data were addressed	
		(d) Cohort study – If applicable, explain how loss to follow-up was addressed	
		Case–control study – If applicable, explain how matching of cases and controls was addressed	
		Cross-sectional study – If applicable, describe analytical methods taking account of sampling strategy	

(continued)

Table 1 (continued)

	Item number	STROBE items	RECORD items
		(e) Describe any sensitivity analyses	
Data access and cleaning methods			RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.
			RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.
Linkage			RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.
Results			
Participants	13	(a) Report the numbers of individuals at each stage of the study (e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed)	RECORD 13.1: Describe in detail the selection of the persons included in the study (i.e., study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.
		(b) Give reasons for nonparticipation at each stage.	
		(c) Consider use of a flow diagram	
Descriptive data	14	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders	
		(b) Indicate the number of participants with missing data for each variable of interest	
		(c) Cohort study – summarise follow-up time (e.g., average and total amount)	
Outcome data	15	*Cohort study* – Report numbers of outcome events or summary measures over Time	
		Case–control study – Report numbers in each exposure category, or summary measures of exposure	

		Cross-sectional study – Report numbers of outcome events or summary measures	
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95 % confidence interval). Make clear which confounders were adjusted for and why they were included	
		(b) Report category boundaries when continuous variables were categorized	
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	
Discussion			
Key results	18	Summarise key results with reference to study objectives	
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.
Interpretation	19	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	
Generalisability	21	Discuss the generalisability (external validity) of the study results	

(continued)

Table 1 (continued)

	Item number	STROBE items	RECORD items
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	
Accessibility of protocol, raw data, and programming code			RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.

Reproduced with permission from Benchimol et al. (2015)

basis for evidence-based medicine. If individuals or groups conducting studies to collate all the evidence on a subject cannot identify, in a systematic manner, studies using routinely-collected health data, then the results of the subsequent systematic review will be skewed, potentially missing important data and results on the basis that they were not easily identifiable from the literature.

Moreover, being able to find studies using routinely-collected health data is a prerequisite to being able to critique and build upon them. Clearly identifying the study as using routinely-collected health data, and the specific datasets and regions, could serve to facilitate necessary replication, reduce unnecessary repetition when relevant studies can be clearly identified, thus reducing inefficient use of resources and potentially facilitating the uptake of relevant study findings. As such, the methodological focus of the guideline will serve to greatly assist those reviewing studies; that is, it will serve as a tool to address the external use of research findings by making these findings more accessible through more complete reporting.

As indicated earlier, the focus of much of the identified guidance related to methodological aspects. Routinely-collected health data has the potential for many sources of data errors or biases, and thus the ability to critique methodological aspects of studies will be important. However, contextual aspects, included the detailing of study population, variables used within the study, and data and linkage, will be important if reproducibility and replication are important goals and we are to avoid a 'cargo cult' approach to research on routinely-collected health data. As before, these aspects facilitate evaluation with respect to the potential applicability of findings to the reader's own context, and enable assessment of whether studies could be replicated if necessary: are similar datasets or variables accessible that would permit replication in order to assess the generalisability or applicability of these findings? In addition, studies that fail to provide such information may be viewed with caution such that the verification of results (through replication) is undertaken in order to prevent the potential for inappropriate uptake of results. Furthermore, the publication of validation algorithms may facilitate knowledge transfer. Studies that have developed and validated algorithms to identify cohorts of patients may be tested in other settings or developed further, contributing to the scientific knowledge base.

Finally, and as indicated within the introduction to this chapter, routinely-collected health data are collected without a specific *a priori* research question in mind and the reasons motivating the data collection may vary. This may affect the quality and applicability of the data to the research questions being examined. In particular, eligibility criteria for inclusion within a database may be important with respect to assessing the applicability of findings. Finally, the authors should be clear as to the extent of their access to the database in question. They should give clear indication as to whether they were provided with a 'cut' of data, or did they have access to all data and conduct their own data cleaning and analyses? These elements represent aspects of individual researcher integrity – an honest presentation of findings that are necessary for developing attributions of truthfulness, but also in terms of assessing the potential impact and relevancy as well as for critical appraisal of the study.

6 Conclusion

Recent years have seen unprecedented advances in data creation, storage and analytics. In particular, there have been great strides in the collection of biomedical data that may be contained in multiple research and healthcare repositories. Healthcare big data sources include routinely-collected data within disease registries, clinical databases such as those in primary care, health administrative data, and electronic health record data. These data offer opportunities to conduct large-scale population research at costs far less than those required for primary data collection and have the potential to provide crucial insights into health services policy and planning. However, they also present important questions that warrant consideration.

We have suggested that a particular area of concern, and one that has to date received too little attention, is the post-analytic reporting of research using biomedical big data. Specifically, there has been a lack of attention paid to the quality of reporting for studies using large-scale health-related datasets. There are important practical reasons for wishing to improve the quality of reporting in studies using routinely-collected health data: the ability to conduct appropriate peer review, to identify areas of need, and to ensure trust in research such that public support is retained and allows the ongoing conduct of research. These practical benefits are derived from important ethical underpinnings that, despite interest in improving the quality of reporting, have been hitherto under-considered. In particular we note that the desire to improve the reporting of research may derive from concerns over the integrity of individual researchers, but may also reflect external drivers that pertain to the use of research results.

The corrective action taken to date has been the development of reporting guidelines. These are often specific to different study designs and reflect the concerns, and views, of experts within a particular field of study. Notably, there has been a lack of guidance for researchers using biomedical big data drawn from routinely-collected health data. We have provided the example of the RECORD Statement as an example of such a reporting guideline.

The development of the RECORD Statement also addresses many of the underlying ethical motivations for the improvement of research reporting. The Delphi process – in which experts in the field come together and engage in a deliberative process of generating and finalising the indicators – is a process that addresses many concerns relating to the individual integrity of researchers: they derive from internal reflections on professional practice, but also external aspects through the promotion of compliance, thus facilitating peer review.

However, the RECORD Statement is only a first step. It will require time for this guidance to permeate into the field of researchers conducting analyses using routinely-collected health data. It also remains an open question as to the effect that the guideline will have on study reporting. Ongoing evaluation will be essential.

It is here where the larger question looms regarding the impact, or potential impact, of reporting guidelines in terms of addressing the underlying ethical concerns that have driven guideline development. Given the nature of reporting

guidelines as a tool for researchers – one that is derived from professional norms and serves as a form of self-regulation – evaluation has tended to focus on intermediary outcomes or the extent to which published studies comply with the elements within guidelines. Even then, there has been surprisingly limited work in terms of evaluation of guidelines (Simera et al. 2008). Indeed, the direct beneficiaries of reporting guidelines are generally researchers and those who peer review research (Altman and Simera 2014).

Work is still required as to the role played by reporting guidelines in the uptake of research, particularly within the realm of health policy. Indeed, there is a paucity of research on the role of reporting guidelines in terms of reducing research waste. In part this is because the endpoints are difficult to establish and are more diffuse in nature, but it is also generally beyond the realm of the researcher. As such, despite the theoretical benefits of reporting guidelines and the data on intermediate endpoints such as completeness of reporting, there remains a dearth of research regarding whether reporting guidelines are addressing many of the external ethical motivators behind reporting guidelines. We see this as an area ripe for further work to advance the evidence-base and improve the quality of reporting for studies using routinely-collected data and other forms of biomedical big data.

References

Altman, D.G., and D. Moher. 2013. Declaration of transparency for each research article. *British Medical Journal* 347: f4796. doi:10.1136/bmj.f4796.

Altman, D.G., and D. Moher. 2014. Importance of transparent reporting of health research. In *Guidelines for reporting health research: A user's manual*, ed. D. Moher, D.G. Altman, K.F. Schulz, I. Simera, and E. Wager, 3–13. Hoboken: Wiley.

Altman, D.G., and I. Simera. 2014. Using reporting guidelines effectively to ensure good reporting of health research. In *Guidelines for reporting health research: A user's manual*, ed. D. Moher, D.G. Altman, K.F. Schulz, I. Simera, and E. Wager, 32–40. Hoboken: Wiley.

Anderson, J.A., M. Eijkholt, and J. Illes. 2013. Ethical reproducibility: Towards transparent reporting in biomedical research. *Nature Methods* 10(9): 843–845. doi:10.1038/nmeth.2564.

Anderson, J.A., B. Sawatzky-Girling, M. McDonald, D. Pullman, R. Saginur, H.A. Sampson, and D.J. Willison. 2011. Research ethics broadly writ: Beyond REB review. *Health Law Review* 19(3): 12–24.

Armstrong, R., E. Waters, L. Moore, E. Riggs, L.G. Cuervo, P. Lumbiganon, and P. Hawe. 2008. Improving the reporting of public health intervention research: Advancing TREND and CONSORT. *Journal of Public Health (Oxford, England)* 30(1): 103–109. doi:10.1093/pubmed/fdm082.

Beauchamp, T.L. 1996. Moral foundations. In *Ethics and epidemiology*, ed. S. Coughlin and T.L. Beauchamp, 24–52. New York: Oxford University Press.

Benchimol, E.I., D.G. Manuel, T. To, A.M. Griffiths, L. Rabeneck, and A. Guttmann. 2011. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *Journal of Clinical Epidemiology* 64(8): 821–829. doi:10.1016/j.jclinepi.2010.10.006.

Benchimol, E.I., L. Smeeth, A. Guttman, K. Harron, D.G. Moher, I. Petersen, H.T. Sørensen, E. von Elm, S.M. Langan, and RECORD Working Committee. 2015. The REporting of studies

Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 12(10): e1001885. doi:10.1371/journal.pmed.1001885.

Bossuyt, P.M., J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L.M. Irwig, J.G. Lijmer, et al. 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal* 326: 41–44.

Butler, D. 2013. When Google got flu wrong. *Nature* 494: 155–156.

Canadian Institutes of Health Research. 2011. *Canada's Strategy for Patient-Oriented Research. Improving health outcomes through evidence-informed care*. Ottawa: Canadian Institutes of Health Research.

Carter, P., G. T. Laurie, and M. Dixon-Woods. 2015. The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*. doi:10.1136/medethics-2014-102374.

Chalmers, I., and P. Glasziou. 2009. Avoidable waste in the production and reporting of research evidence. *Lancet* 374: 86–89. doi:10.1016/s01406736(09)60329-9.

Cohen, I.G., R. Amarasingham, A. Shah, B. Xie, and B. Lo. 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs* 33(7): 1139–1147. doi:10.1377/hlthaff.2014.0048.

Collins, F.S., and L.A. Tabak. 2014. Policy: NIH plans to enhance reproducibility. *Nature* 505(7485): 612–613.

Cook, J.A., and G.S. Collins. 2015. The rise of big clinical databases. *British Journal of Surgery* 102(2): e93–e101. doi:10.1002/bjs.9723.

Cook, S., C. Conrad, A.L. Fowlkes, and M.H. Mohebbi. 2011. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One* 6(8), e23610. doi:10.1371/journal.pone.0023610.t001.

Copeland, P., R. Romano, T. Zhang, G. Hecht, D. Zigmond, and C. Stefansen. 2013. Google disease trends: An update. Paper presented at the International Society of Neglected Tropical Disease (ISNTD) Bites, London, October 17, 2012.

Council of Canadian Academies. 2015. Accessing health and health-related data in Canada. Ottawa (ON): The expert panel on timely access to health and social data for health research and health System innovation, Council of Canadian Academies.

Currie, J. 2013. "Big Data" versus "Big Brother": On the appropriate use of large-scale data collections in pediatrics. *Pediatrics* 131(Suppl 2): S127–S132. doi:10.1542/peds.2013-0252c.

Damschroder, L.J., J.L. Pritts, M.A. Neblo, R.J. Kalarickal, J.W. Creswell, and R.A. Hayward. 2007. Patients, privacy and trust: Patients' willingness to allow researchers to access their medical records. *Social Science and Medicine* 64(1): 223–235. doi:10.1016/j.socscimed.2006.08.045.

Davenport, P., W. Cragg, M. Crago, D. Fanelli, J.-M. Fleury, L.M. Given, R. Heslegrave, et al. 2010. *Honesty, accountability and trust: Fostering research integrity in Canada*. Ottawa: Council of Canadian Academies.

de Lusignan, Simon, and Chris van Weel. 2006. The use of routinely-collected computer data for research in primary care: Opportunities and challenges. *Family Practice* 23(2): 253–263.

Dean, Bonnie B., Lam Jessica, Jaime L. Natoli, Butler Qiana, Aguilar Daniel, and Robert J. Nordyke. 2009. Review: Use of electronic medical records for health outcomes research A literature review. *Medical Care Research and Review* 66(6): 611–638.

Elliott, K.C., and D.B. Resnik. 2014. Science, policy, and the transparency of values. *Environmental Health Perspectives* 122(7): 647–650. doi:10.1289/ehp.1408107.

Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network. Essential resources for writing and publishing health research. http://www.equator-network.org/. Accessed 21 Sept 2015.

Feynman, R.P. 1998. Cargo cult science. In *The art and science of analog circuit design*, ed. J. Williams. Wolburn: Butterworth-Heinemann.

Franklin, S.S., M.G. Larson, S.A. Khan, N.D. Wong, E.P. Leip, W.B. Kannel, and D. Levy. 2001. Does the relation of blood pressure to coronary heart disease risk with aging? The Framingham Heart study. *Circulation* 103: 1245–1249.

Friedman, S.L. 2007. Finding treasure: Data sharing and secondary analysis in developmental science. *Journal of Applied Developmental Psychology* 28(5-6): 384–389. doi:10.1016/j.appdev.2007.07.001.

Glasziou, Paul, Douglas G. Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, and Elizabeth Wager. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet* 383(9913): 267–276. doi:10.1016/s0140-6736(13)62228-x.

Groves, T. 2008. Enhancing the quality and transparency of health research. *British Medical Journal* 337: a718. doi:10.1136/bmj.a718.

Guttmann, A., M. Nakhla, M. Henderson, T. To, D. Daneman, K. Cauch-Dudek, X. Wang, K. Lam, and J. Hux. 2010. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. *Pediatric Diabetes* 11(2): 122–128. doi:10.1111/j.1399-5448.2009.00539.x.

Harpe, Spencer E. 2009. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 29(2): 138–153.

Hemkens L.G., E.I. Benchimol, S.M. Langan, M. Briel, B. Kasenda , J.M. Januel , E. Herrett, and E. von Elm. The reporting of studies using routinely collected health data is often insufficient: Systematic literature analysis. *Journal of Clinical Epidemiology*, in press.

Hoffman, S., and A. Podgurski. 2013. The use and misuse of biomedical data: Is bigger really better? *American Journal of Law and Medicine* 39: 497–538.

Institute of Medicine. 2015. *Sharing clinical trial data: Maximizing benefits, minimizing risk*. Washington, D.C.: Institute of Medicine.

Ioannidis, J.P.A. 2012. The importance of potential studies that have not existed and registration of observational data sets. *JAMA* 308(6): 575–576.

Jacobs, J.P., R.J. Cerfolio, and R.M. Sade. 2009. The ethics of transparency: Publication of cardiothoracic surgical outcomes in the lay press. *Annals of Thoracic Surgery* 87(3): 679–686. doi:10.1016/j.athoracsur.2008.12.043.

Kilkenny, C., W.J. Browne, I.C. Cuthill, M. Emerson, and D.G. Altman. 2010. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biology* 8(6): e1000412. doi:10.1371/journal.pbio.1000412.t001.

King, G. 2016. Preface: Big data is not about the data. In *Computational social science: Discovery and prediction*, ed. R.M. Alvarez. Cambridge: Cambridge University Press.

Kreuter, F., and R.D. Peng. 2014. Extracting information from Big Data: Issues of measurement, inference and linkage. In *Privacy, big data, and the public good. Frameworks for engagement*, ed. J. Lane, S. Stodden, S. Bender, and H. Nissenbaum, 11–20. Cambridge: Cambridge University Press.

Lane, Julia, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. 2014. Editors' Introduction. In *Privacy, big data, and the public good. Frameworks for engagement*, ed. J. Lane, S. Stodden, S. Bender, and H. Nissenbaum, 11–20. Cambridge: Cambridge University Press.

Langan, S.M., E.I. Benchimol, A. Guttmann, D. Moher, I. Petersen, L. Smeeth, H.T. Sorensen, F. Stanley, and E. Von Elm. 2013. Setting the RECORD straight: Developing a guideline for the REporting of studies Conducted using Observational Routinely-collected Data. *Clinical Epidemiology* 5: 29–31. doi:10.2147/CLEP.S36885.

Last, J. 1996. Professional standards of conduct for epidemiologists. In *Ethics and epidemiology*, ed. S. Coughlin and T.L. Beauchamp, 53–75. New York: Oxford University Press.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014a. Big data. The parable of Google Flu: Traps in big data analysis. *Science* 343(6176): 1203–1205. doi:10.1126/science.1248506.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014b. Google flu trends still appears sick: An evaluation of the 2013–2014 flu season. Available at SSRN 2408560.

Maeng, M., H.H. Tilsted, L.O. Jensen, L.R. Krusell, A. Kaltoft, H. Kelbaek, A.B. Villadsen, et al. 2014. Differential clinical outcomes after 1 year versus 5 years in a randomised comparison of zotarolimus-eluting and sirolimus-eluting coronary stents (the SORT OUT III study): A multicentre, open-label, randomised superiority trial. *Lancet* 383(9934): 2047–2056. doi:10.1016/s0140-6736(14)60405-0.

Manson, N.C. 2010. Why do patients want information if not to take part in decision making? *Journal of Medical Ethics* 36(12): 834–837. doi:10.1136/jme.2010.036491.

Manuel, D.G., L.C. Rosella, and T.A. Stukel. 2010. Importance of accurately identifying disease in studies using electronic health records. *British Medical Journal* 341: c4226. doi:10.1136/bmj.c4226.

Masic, I. 2012. Ethical aspects and dilemmas of preparing, writing and publishing of the scientific papers in the biomedical journals. *Acta Informatica Medica* 20(3): 141–148. doi:10.5455/aim.2012.20.141-148.

Mittelstadt, B.D., and L. Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.

Moher, D. 2007. Reporting research results: A moral obligation for all researchers. *Canadian Journal of Anesthesia* 54(5): 331–335.

Moher, D., D.J. Cook, S. Eastwood, I. Olkin, D. Rennie, and D.F. Stroup. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 354(9193): 1896–1900.

Moher, D., I. Simera, K.F. Schulz, J. Hoey, and D.G. Altman. 2008. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Medicine* 6: 13. doi:10.1186/1741-7015-6-13.

Moher, D., A. Liberati, J. Tetzlaff, D.G. Altman, and The PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6(7): e1000097. doi:10.1371/.

Moher, D., K.F. Schulz, I. Simera, and D.G. Altman. 2010. Guidance for developers of health research reporting guidelines. *PLoS Medicine* 7(2): e1000217. doi:10.1371/journal.pmed.1000217.t001.

Needleman, I., D. Moher, D.G. Altman, K.F. Schulz, D.R. Moles, and H. Worthington. 2008. Improving the clarity and transparency of reporting health research: A shared obligation and responsibility. *Journal of Dental Research* 87(10): 894–895.

Nicholls, S.G., P. Quach, E. von Elm, A. Guttmann, D. Moher, I. Petersen, H.T. Sørensen, L. Smeeth, S.M. Langan, and E.I. Benchimol. 2015. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. *PLoS One* 10(5): e0125620. doi:10.1371/journal.pone.0125620. 10.5061/dryad.7d65n.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*. London: Nuffield Council on Bioethics.

O'Connor, A. 2010. Reporting guidelines for primary research: Saying what you did. *Preventive Veterinary Medicine* 97(3-4): 144–149. doi:10.1016/j.prevetmed.2010.09.010.

O'Neill, O. 2002. *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

Prady, S.L., S.J. Richmond, V.M. Morton, and H. Macpherson. 2008. A systematic evaluation of the impact of STRICTA and CONSORT recommendations on quality of reporting for acupuncture trials. *PLoS One* 3(2): e1577. doi:10.1371/journal.pone.0001577.

Sciencewise. 2014. Big Data. Public views on the collection, sharing and use of personal data by government and companies: Sciencewise Expert Resource Centre, London: UK. http://www.sciencewiseerc.org.uk/cms/assets/Uploads/SocialIntelligenceBigData.pdf.

Simera, I., D.G. Altman, D. Moher, K.F. Schulz, and J. Hoey. 2008. Guidelines for reporting health research: The EQUATOR network's survey of guideline developers. *PLoS Medicine* 5(6): e139. doi:10.1371/journal.pmed.0050139.t001.

Simera, I., D. Moher, A. Hirst, J. Hoey, K.F. Schulz, and D.G. Altman. 2010a. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Medicine* 8: 24. doi:10.1186/1741-7015-8-24.

Simera, I., D. Moher, J. Hoey, K.F. Schulz, and D.G. Altman. 2010b. A catalogue of reporting guidelines for health research. *European Journal of Clinical Investigation* 40(1): 35–53. doi:10.1111/j.1365-2362.2009.02234.x.

Sorensen, A.A., R.D. Wojahn, M.C. Manske, and R.P. Calfee. 2013. Using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement to assess reporting of observational trials in hand surgery. *Journal of Hand Surgery. American Volume* 38(8): 1584–1589. doi:10.1016/j.jhsa.2013.05.008. e1582.

Taylor, Mark. 2014. Information governance as a force for good? Lessons to be learnt from care.data. *SCRIPTed* 11(1). doi:10.2966/scrip.110114.1. https://scripted.org/article/information-governance-force-good-lessons-learnt-care-data/.

ter Meulen, R.H.J., A.J. Newson, M.-R. Kennedy, and B. Schofield. 2011. *Background paper: Genomics, health records, database linkage and privacy*. London: Nuffield Council on Bioethics.

Thompson, P.A., and A. Burnett. 2012. *Reproducible research. CORE Issues in Professional and Research Ethics*, Issue 1, Paper 6. https://nationalethicscenter.org/resources/734/download/Thompson.pdf.

Turner, L., L. Shamseer, D. G. Altman, L. Weeks, J. Peters, T. Kober, S. Dias, K. F. Schulz, A. C. Plint, and D. Moher. 2012. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews* 11:Mr000030. doi:10.1002/14651858.MR000030.pub2.

Vasilevsky, N.A., M.H. Brush, H. Paddock, L. Ponting, S.J. Tripathy, G.M. Larocca, and M.A. Haendel. 2013. On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ* 1: e148. doi:10.7717/peerj.148.

Vayena, E., M. Salathe, L.C. Madoff, and J.S. Brownstein. 2015. Ethical challenges of big data in public health. *PLoS Computational Biology* 11(2): e1003904. doi:10.1371/journal.pcbi.1003904.

von Elm, E., D.G. Altman, M. Egger, S.J. Pocock, P.C. Gotzsche, and J.P. Vandenbroucke. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Epidemiology* 18(6): 800–804. doi:10.1097/EDE.0b013e3181577654.

Wager, E., and S. Kleinert. 2011. Responsible research publication: International standards for authors. A position statement at the 2nd World Conference on Research Integrity, Singapore, July 22–24, 2010. In *Promoting research integrity in a global environment*, ed. T. Mayer and N. Steneck, 309–316. Singapore: Imperial College Press/World Scientific Publishing.

Wang, W., and E. Krishnan. 2014. Big data and clinicians: A review on the state of the science. *JMIR Med Inform* 2(1): e1. doi:10.2196/medinform.2913.

Weiskopf, N.G., and C. Weng. 2013. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association* 20(1): 144–151. doi:10.1136/amiajnl-2011-000681.

World Medical Association. 2013. *WMA declaration of Helsinki – Ethical principles for medical research involving human subjects*. Fortaleza: World Medical Association.

Creating a Culture of Ethics in Biomedical Big Data: Adapting 'Guidelines for Professional Practice' to Promote Ethical Use and Research Practice

Rochelle E. Tractenberg

Abstract Two scientific domains that are crucial in "Biomedical Big Data", computing and statistics, do not typically require "training in the responsible conduct of research" or research ethics. While "responsible conduct of research" (RCR) comprises interactions with subjects (human and non-human), it also involves interactions with other scientists, the scientific community, the public, and in some contexts, research funders. Historically, the development or emergence of disciplines and professions tend to involve a semi-simultaneous emergence of professional norms and/or codes of conduct. However, Biomedical Big Data is not emerging as a single discipline or profession, and engages practitioners from many diverse backgrounds. Moreover, the place of the data analyst or the computer scientist developing analytic algorithms seems to be too granular to be considered specifically within the activities that comprise "responsible research and innovation" (RRI). Current legal and policy-level considerations of Biomedical Big Data and RRI are implicitly assuming that scientists carrying out the research and achieving the innovations are exercising their scientific freedom – i.e., conducting research – responsibly. The assumption is that all scientists are trained to conduct research responsibly. In the United States, federal agencies funding research require that training in RCR be included – some of the time. Because the vast majority of research that was federally funded has *not* included Biomedical Big Data, RCR training paradigms have emerged over the past 20 years in US institutions that are not particularly relevant for Big Data. While it would be efficient to utilize such established, well-known, easily-documented RCR training programs, this chapter discusses how and why this is less likely to support the development of professional norms that are relevant for Biomedical Big Data. This chapter will describe an alternative approach that can support ongoing reflection on professional obligations, which can be used in a wide range of ethical, legal, and social implications (ELSI),

R.E. Tractenberg, PhD, MPH, PhD, PStat®, FASA (✉)
Director, Collaborative for Research on Outcomes and – Metrics, Departments of Neurology, Biostatistics, Bioinformatics & Biomathematics, and Rehabilitation Medicine, Georgetown University Medical Center, Building D, Suite 207, 4000 Reservoir Rd. NW, Washington, DC, USA
e-mail: rochelle.tractenberg@gmail.com

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_16

including those that have not yet been identified. This may be the greatest strength of this alternative approach for preparing practitioners for Biomedical Big Data, because the ability to apply prior learning in ethics to previously unseen problems is especially critical in the current era of dynamic and massive data accumulation. To support the development of normative ethical practices among practitioners in Biomedical Big Data, this chapter reviews the guidelines for professional practice from three statistical associations (American Statistical Association; Royal Statistics Society; International Statistics Institute) and from the Association of Computing Machinery. These can be leveraged to ensure that, in their work with Biomedical Big Data, participants know and understand the ethical, legal, and social implications of that work. Formal integration of these (or other relevant) guidelines into the preparation for practice with data (big and small) can help in dealing with ethical challenges *currently* arising with Big Data in biomedical research; moreover, this integration can also help deal with challenges that have not yet arisen. These outcomes, which are consistent with recent calls for the institutionalization of reflection and reasoning around ELSI across scientific disciplines, in Europe, are only possible as long as the integration effort does *not* follow a currently-dominant paradigm for training in RCR. Preparing scientists to engage competently in conversations around ethical issues in Biomedical Big Data requires purposeful, discipline-relevant, and developmental training that can come from, and support, a culture of ethical biomedical research and practice with Big Data.

1 Introduction

> *Both scholarship and teaching in any field reflect the character of inquiry, the nature of community, and the ways in which research and teaching are conducted in that particular discipline or disciplinary intersection*. (Shulman 2008, p. xii).

Big Data can be defined in a variety of manners, relating to the amount of data; speed with which it is accumulated; complexity of the data and/or the interrelationships it contains; or some combination of these characteristics. An alternative perspective, articulated by Boyd and Crawford (2012) is that "Big Data" – capitalized to represent its status as a cultural phenomenon (p. 675) – differs from earlier massive-scale data sets (e.g., national medical data registries; national census databases) in its potential or requirement for the integration of multiple and diverse domains of expertise. The construct of "ethics in Biomedical Big Data" might refer to the ethical collection, management, or use of this data; it might also refer to the normative behaviours that guide or define professional practice and activities utilizing this type of data. Two interesting perspectives on the potential for misuse of "Big Data", which reinforce this construct of "ethics in Biomedical Big Data", come from Owen et al. (2012), where a focus on promoting "responsible research and innovation" tends to emphasize the realization of social good (through research and innovation); and from Elias (2014), where the focus is on the protection of privacy for individuals whose micro data are combined from within and across European

Union nations in to "Big Data". Steinmann et al (2015) elaborate very concretely how challenging privacy can be to define – and protect- in the modern era, and Elias (2014) discusses legislation in the United Kingdom that targets the protection of privacy while facilitating Big Data-driven research. Owen et al. (2012) describe broader, more social (less private) considerations than those on which the European model of Responsible Research and Innovation (RRI) seems to be focused. That is, current high-level discussions in the European context may be focused on policy and law whereby societal good can be achieved with research and innovation that involves Big Data; much of the discussion is focused on privacy. The place of the data analyst or the computer scientist developing analytic algorithms seems to be too granular to be considered specifically within the activities that comprise RRI, which is not surprising since their contributions to RRI may seem to be more supportive than directive. That is, the "science" that supports RRI may depend quite critically on RRI within statistics or computer science, but it is not those innovations that the discussion around RRI and its ethical considerations tends to revolve.

An additional complication is that current legal and policy-level considerations of Biomedical Big Data and RRI are implicitly assuming that all of the scientists carrying out the research and achieving the innovations are exercising their scientific freedom – i.e., conducting research – *responsibly*. In fact, the assumption is that all scientists are *trained* to conduct research responsibly. In the United States, federal agencies funding research require that training in RCR be included – some of the time. Because the vast majority of research that was federally funded has *not* included Biomedical Big Data, RCR training paradigms have emerged over the past 20 years in US institutions that are not particularly relevant for Big Data.

Owen et al. (2012) note that "(s)cientists already have responsibilities, including those associated with the concepts of research integrity . . . RRI, however, confers new responsibilities." (p. 756). It has been argued (Tractenberg et al. 2015) that practitioners who are being trained for a career in or with Biomedical Big Data (BBD), whether in or outside of academia, need specific training to appropriately reason through challenges arising from the ethical, legal, and social issues (ELSI) their engagement with BBD may entail. While some applications of data mining are not considered "research" *per se*, responsible and professional practice with data can be argued to pertain, and so "ethics in BBD" can be conceptualized as *responsible conduct in/with BBD*. As Shulman articulated, any scientific field reflects the "ways in which research and teaching are conducted in that particular discipline or disciplinary intersection" (2008, p. xii). Since "Biomedical Big Data" is itself a "disciplinary intersection", contemplation of ethics in BBD must include consideration of biomedical research domains (e.g., clinical, microbiological, or genetic), data collection and management (across technological domains), and analytic (statistics, biostatistics) domains. Responsible conduct must be considered a core curricular feature for those who are being trained for a career that involves BBD –because ethical norms can support the development of sense of professionalism/professional identity (Tractenberg et al. 2015); in their discussion of RRI, Owen et al. (2012) as much as assume that this is actually true (p. 755–756). One feature of the European discussion around RRI that has been less well explored in the US context is the

role of the citizenry in either establishing the priorities for US funders to target in research (or innovation); if *all* students received training around ELSI – and how to reason about them – in Big Data, perhaps the conversation in the US might reach a level of social engagement that the European RRI discussions appear to have (see Owen et al. 2012 for discussion of the variety of programs and policy discussions in Europe around RRI). The objective of integrating ethical reasoning into all college-level training may be less achievable than that of training *all* those who receive federal funding to develop their scientific skills in the responsible conduct of their research, whatever their contribution to that research might be. However, even if students are not being prepared to engage in research *per se*, the initiation of a sense of professional identity can be seeded with a focus on aspects of professional, ethical, practice with data (big or small, biomedical or otherwise). This is consistent with the conversations around RRI in Europe (Owen et al. 2012) and the United Kingdom (Elias 2014), and as discussed in the next sections, *not* consistent with the dominant RCR training paradigm in the US.

If educational institutions will continue the current trend of developing new degree or certificate programs to prepare students technically to engage with BBD, the promotion of acceptable, ethical, behavioural norms for these future practitioners should also be considered. It would be efficient in some sense to utilize established, well-known, easily-documented RCR training programs; however, in this chapter it is argued that BBD as an emerging domain or discipline would be better served if new and relevant training, emphasizing reasoning and professionalism, were developed instead. One paradigm for ethics training in biomedical research, currently dominant across the US, is discussed and its weaknesses are articulated. A new paradigm for training and promoting common values and behavioural norms for ethical practice around BBD is described, together with potentially useable or adaptable materials.

2 Training in "Responsible Conduct of Research"

A typical conceptualization of ethics education in biomedical research in the United States (US) is training in the "responsible conduct of research" (RCR). Graduate-level science instruction might include one module or course –to be completed at least once per 4 years (National Institutes of Health 2009). For biomedical researchers in the United States, research funding is (or has historically been) obtained primarily through grants from the National Institutes of Health (NIH), and the NIH regulations about RCR training have created the dominant training paradigm around RCR. Examining this RCR training paradigm as representative of "ways in which research and teaching *ethics* are conducted in that particular discipline or disciplinary intersection" (Shulman 2008; "ethics" and emphasis added) has been very informative for our own research program into teaching and learning in research ethics. Specifically we reject the NIH RCR training paradigm because:

1. Not all scientists trained using federal funds **must** receive training in the responsible conduct of research.
2. There are no requirements to document the capacity to reliably train others in RCR or research ethics.
3. RCR training that *does* follow the NIH RCR training paradigm actually *does not* work.

The lesson to be learned from our extensive engagement with this paradigm is that it should no longer be the dominant RCR training paradigm. We elaborate in the following sections how these three observations should inform those seeking to create (or identify) training opportunities around ethics and professional practice with BBD – and should lead to the conclusion that the NIH RCR training paradigm should *not* be adopted by new disciplines to "reflect the character of inquiry, the nature of community, and the ways in which research and teaching are conducted" (to paraphrase Shulman 2008).

Not all scientists trained using federal funds **must** receive training in the responsible conduct of research. The National Institutes of Health (NIH) outlined new rules that scientists proposing to train new scientists, or trainees seeking fellowship training with federal funds, "must" document their RCR training plans (NIH 2009). However, also according to the NIH, not all researchers who will receive NIH funded training in research must also receive RCR training: only "(i)ndividuals who will be involved in the design or conduct of NIH-funded human subjects research must fulfil the education requirement." (http://grants.nih.gov/grants/policy/hs_educ_faq.htm#132). The definition of "human subjects research" is, "(r)esearch is considered to involve human subjects when an investigator conducting research obtains (1) data through intervention or interaction or with a living individual, or (2) identifiable private information about a living individual." (Department of Health and Human Subjects 2009). Specifically for most Biomedical Big Data applications, this *excludes* "(r)esearch involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects." (Department of Health and Human Subjects 2009). The FAQ continues, "(i)nvestigators who conduct studies with human specimens, tissues, or data that are determined not to involve human subjects are not required to fulfil the education requirement." This caveat means that only some individuals who are trained in science using public funding must also be trained in the responsible conduct of research.

This is particularly unreasonable given that the list of topics that the NIH considers critical for RCR training (NIH 2009) is actually quite extensive and contains important information for competent scientific practice that is actually independent of the specific type (or source) of data a scientist utilizes:

- conflict of interest – personal, professional, and financial
- policies regarding human subjects, live vertebrate animal subjects in research, and safe laboratory practices

- mentor/mentee responsibilities and relationships
- collaborative research including collaborations with industry
- peer review
- data acquisition and laboratory tools; management, sharing and ownership
- research misconduct and policies for handling misconduct
- responsible authorship and publication
- the scientist as a responsible member of society, contemporary ethical issues in (biomedical) research, and the environmental and societal impacts of scientific research.

If a program or a trainee can make an argument that their research does not qualify as "human subjects", then their omission of any training in the topics listed above is supported by the NIH. This is contrary to the original NIH Policy published in 1989, which stated that, "(e)ffective July 1, 1990, all competing National Research Service Award institutional training grant applications must include a description of the formal or informal activities related to the instruction about the responsible conduct of research that will be incorporated into the proposed research training program." (NIH 1989; p. 1). However, 20 years later, the NIH deems "responsible conduct in research" to be outside of the educational requirements of some trainees who will receive public funding for their training and research. Following this paradigm, students trained in Big Data outside of biomedical centres or contexts require no particular preparation prior to applying that training and knowledge to Biomedical Big Data. The core features of "professional practice" (articulated for computer science and statistics in later sections) are highly relevant for Big Data in and outside of biomedical contexts, but NIH considers these to be insufficient preparation – but only for those who will use "human subjects". An implication is that they might represent too much preparation for those whose work does not meet the "human subjects" definition.

The implication of the NIH paradigm, that *only* scientists using "human subjects" in their research need training for responsible conduct, is inconsistent with the background for the recent restated requirement for this training. NIH (2009) explicitly points out that the peer review of RCR training plans should be guided by the "Basic Principles" of the responsible conduct of research (see below). The six Basic Principles articulated are:

1. Responsible conduct of research is an essential component of research training. Therefore, instruction in responsible conduct of research is an integral part of *all research training programs*, and its evaluation will impact funding decisions. *(Emphasis added.)*
2. Active involvement in the issues of responsible conduct of research should occur throughout a scientist's career. Instruction in responsible conduct of research should therefore be appropriate to the career stage of the individuals receiving training.
3. Individuals supported by individual funding opportunities such as fellowships and career development awards are encouraged to assume individual and personal responsibility for their instruction in responsible conduct of research.

4. Research faculty of the institution should participate in instruction in responsible conduct of research in ways that allow them to serve as effective role models for their trainees, fellows, and scholars.
5. Instruction should include face-to-face discussions by course participants and faculty; i.e., on-line instruction may be a component of instruction in responsible conduct of research but is not sufficient to meet the NIH requirement for such instruction, except in special or unusual circumstances.
6. Instruction in responsible conduct of research must be carefully evaluated in all NIH grant applications for which it is a required component (*i.e., not for applications describing research that does not meet the "human subjects" definition).*

The fact that this training, and these basic principles, are only actually required for – and therefore conceptualized as relevant to- scientists who deal with "human subjects data" reflects an absurd circularity. The six "Basic Principles" and indeed the topics to be covered in "RCR training" are extremely important; however the NIH RCR training *paradigm* is not consistent with the principles, and the paradigm tends to undermine the importance of the topics by suggesting that they are not relevant for all scientists. Any training paradigm that is adopted for programs preparing students to engage with BBD should be consistently applied, and the applicability of the training should be clear to all involved. If the NIH RCR training paradigm was to be accepted for BBD, it would mean that no students would receive training to promote responsible conduct in science or professional practice, and while that might seem efficient, the *applicability* of that approach for a future of ethical work with BBD is arguably negligible.

Another reason not to adopt the NIH RCR training paradigm for a new field or disciplinary intersection like BBD is that there are no requirements to document the capacity to reliably train others in RCR or research ethics. The 2009 NIH rules stipulate that practicing scientists must complete "training" not less than once in 4 years, and that claims of competence as trainers or mentors in the responsible conduct of research can be supported by leading other RCR training activities. However, there is no requirement that the training vary, or change in sophistication, according to the career stage of the practicing scientist; in some cases the exact same online modules (e.g., CITI training) can or must be completed by students, staff, and faculty alike to satisfy this requirement. There is no mechanism for learning how to train others; in fact, there is no difference in the training of faculty, staff or students in RCR training within any given institution (that can be found online – possibly individual programs, research groups, or labs do have varied RCR training opportunities). In general, then, the NIH RCR training paradigm tends to result in everyone completing the same, static, materials as representing "being trained". This systematic emphasis on simple, compliance-oriented completion of "RCR training" promotes simple, compliance-oriented RCR training. By contrast, training to promote ethical practice in/with BBD can be tailored to also promote a sense of professional identity for those coming from different disciplines. Rather than adopt the institution's existing "RCR training", which might or might not follow

the NIH paradigm, new BBD programs should strive to include a focus on ethics, or on training BBD practitioners to reason through new and emerging ethical, legal, and social issues (ELSI) for their field or their field's contributions to work in BBD. In that model, those with greater responsibilities or who are further along a career trajectory could and would be able to demonstrate their greater sophistication, serving as legitimate role models of the profession, and making the ethics or RCR training both relevant to actual practice and *dynamic*.

There is a great deal of evidence that RCR training that follows the NIH RCR training paradigm does not work. Many reports have been published describing the failure of traditional "RCR training" to produce positive or lasting effects (see e.g. Kalichman and Plemmons 2007; Antes et al. 2009; Schmaling and Blume 2009; Antes et al. 2010; see also Mumford et al. 2009; Fanelli 2009). In their review of the varieties of approaches to required ethics training, Antes et al. (2009) articulate how even within one course, a variety of educational and ethics perspectives may be integrated – whether purposefully or not – and this variety itself may contribute to challenge or undermine the effectiveness of the resulting course. Given this review, and from the perspective of cognitive psychological theory on adult learning/learners (e.g., Knowles and Holton 2005), the only truly surprising aspect of results showing that typical "RCR training" does not work is that it nevertheless remains the dominant paradigm. However, it is precisely this domination in the face of a lack of efficacy that motivated this chapter: alternative RCR training paradigms are difficult to create. If similar exemptions to the NIH's for research that does not fit its "human subjects" definition exist at other institutions, there is no incentive to create new paradigms, and if any RCR training is "required", the existing institutional training can be used. Our research has focused on undermining the dominance of the NIH RCR training paradigm even when the topic (e.g., Biomedical Big Data) does meet the human subjects definition; we also hope to promote consideration of alternative paradigms to support professional norms and ethics around BBD. Currently, the assertion by Owen et al. (2012) that "(s)cientists already have responsibilities, including those associated with the concepts of research integrity . . . RRI, however, confers new responsibilities." (p. 756) is simply not supported for scientists in the US whose research is deemed not to involve "human subjects".

In and outside of BBD, there are ever-increasing numbers of new scientific techniques, and ways to archive, access, and analyse data. The potential for ethical dilemmas to arise simply from the pressures to "succeed" given dropping grant funding levels and other features of competition across the sciences increases as well. Given this dynamic and competitive backdrop, requiring coursework on a static list of topics, with which *only some* scientists should be "familiar" is unlikely to improve responsibility in research. In fact, it has not: since the requirement for RCR training in NIH funding was introduced in the 1989 (for applications submitted in 1990), the rates of falsification, fraud, and plagiarism have increased (see Fang et al. 2012). If publishing a topics list sufficed to inculcate a sense of responsibility in

the conduct of scientific research, NIH and other federal and international funding agencies would not continue to designate funds for research on how to improve ethics education and, since this began over two decades ago, the fraud, falsification and plagiarism rates in published science would have plateaued, and not increased by a factor of 10 since 1975 (Fang et al. 2012). It is the NIH RCR training paradigm, and not the specific content, that has been shown to fail across many different studies. Table 1 below (reprinted with permission from Tractenberg and FitzGerald 2012) describes the totality of "RCR training opportunities" offered (as of 2011) at their university. Table 1 is notable for two reasons: (A) all but two of the opportunities shown are passive; and (B) there is no way for these opportunities to promote growth in the understanding of what "responsible conduct" might mean with actually increasing responsibility, increasing sophistication, or changes in roles (student, post-doc, junior faculty, senior faculty) throughout a scientific career. One feature of RRI that Owen et al. (2012) identified as "emerging" as of 2012 is the ". . . integration and institutionalization of established mechanisms of reflection . . . around the process of research and innovation." (p. 755).

Clearly, the NIH RCR training paradigm is wholly inadequate to promote mechanisms of reflection around ELSI, because reasoning and reflecting are never mentioned in the typical NIH RCR training opportunity. Moreover, without all scientists participating in active and dynamic RCR training, the "institutionalization" of reflection about the ELSI of research will never be feasible.

Another aspect of the opportunities for "RCR training" in Table 1 is that they are focused on human subjects; not surprising given that this is the only type of investigator or trainee who is required to complete any RCR training. As such, they have little relevance (perceived or actual) for individuals who engage in research using human samples and tissues but that do not "qualify" as "human subjects research". There is no mention of guidance from professional associations- which may be more relevant for investigators using Biomedical Big Data, nor is there any indication that other members of a research team might actually utilize ethical guidelines for professional practice of their discipline to guide decision-making. This is an essential point: Table 1 represents the *institutional* RCR training program- which is NIH compliant- so any new degree program or training grant proposal need only refer to this institutional program to also be compliant. This general, generic approach also implies that *no* training program has unique requirements for ethical practice; it is quite explicit that no level of training has any characteristic that might require more advanced preparation. And while it is very difficult to articulate specific requirements for training that would support ethical practice with BBD, it is not difficult to see that a generic topics list – while it might do much to promote a general sense of what professional behaviour comprises across all scientific domains (if only it were required for all science trainees and practitioners) – is unlikely to prepare new practitioners to handle challenges from ELSI that arise from work in and with BBD.

Table 1 Aligning the Georgetown University Responsible Conduct of Research (RCR) Knowledge, Skills and Abilities (KSAs) and Georgetown University RCR training opportunities appropriate for: 1 = undergraduate; 2 = graduate and pre-doctoral; 3 = post-doctoral; 4 = faculty (F2F = face to face interactions)

Knowledge, Skills and Abilities (KSAs): Opportunities:	Prerequisite knowledge	Recognize a moral issue	Identify decision-making frameworks	Identify and evaluate alternative actions	Make and justify decision	Reflect on decision
Semester-long course on RCR (GUMC Phar 534)	1–2 completion of course	1–2 completion of course	1–2 completion of course	1–2 completion of course	1–2 completion of course	1–2 completion of course
		3–4: not available	3–4: not available	3–4: not available	3–4: not available	3–4: not available
Semester-long course on scientific skills and ethical questions (GUMC NSCI 532)	1–2 completion of course					
	3–4: not available					
Electronic Collaborative Institutional Training Initiative (CITI) RCR course	1–4 completion of course					
National Institutes of Health (NIH) electronic human subject protection training	1–4 completion of course					
Online Health Insurance Portability and Accountability Act (HIPAA) training	1–4 completion of course					
American Association for Laboratory Animal Science (AALAS) modules and training in proper animal research policies and techniques	1–4 completion of course					

New faculty orientation, including RCR info.	1–3 not available 4: attending session					
Mentor/Mentee Discussion	1–4: attending session					
Clinical Research Network seminar	1–4: attending session					
Campus-wide local and/or national programs	1–4: attending session					
Office of Technology Commercialization programs on Invention 2 Innovation	1–4: attending session					
Distribution of the poster entitled "What you need to Know about Ethical Issues when Writing a Scientific Paper"	1–4 receipt/perusal					
Distribution of the University Code of Ethical Conduct	1–4 receipt/perusal					

Adapted from Table 1 from Tractenberg and FitzGerald (2012). Reprinted by permission of Taylor & Francis Ltd, www.tandfonline.com

3 Ethical Principles in Biomedical Research

As outlined above, in the United States, biomedical researchers who propose, or will be trained, to utilize human subjects data using NIH (or any federal) grant funds must complete some RCR training at least once every 4 years. The ethical principles most often invoked in "biomedical research", deriving from the Belmont Report, tend to focus on the behaviour of those in the research enterprise who are in closest contact with human research participants. The ethical principles that are most commonly emphasized in research involving human participants are: respect for persons; beneficence; and justice. These principles, outlined in the seminal book by Beauchamp and Childress (1983), are currently most widely taught and discussed within the main topics of training in "responsible conduct of research" articulated by the NIH. Because few biomedical researchers are aware of the Ethical Guidelines of other disciplines (existence or content), or because they might consider themselves "trained" in the responsible conduct of research once they have completed their institutional training (because the institution – and grant agencies – consider them "trained"), they may never seek to understand how "professional practice" is conceptualized within the other disciplines represented on their research teams. As described earlier, those whose research or practice did not meet the NIH definition of "human subjects" –including experts in data mining, management, and analysis – may initiate or be recruited into BBD projects, without having had *any* training in RCR. Because of its status as dominant (in the US), compounded with the "human subject"-centricity of most NIH-compliant training, the NIH RCR training paradigm would be clearly irrelevant for those new to BBD. The simplest solution is to note that BBD will nearly always fail the NIH "human subjects" definition, and thus there are no considerations or concerns about ethical conduct of BBD projects. There are many cultural forces operating to maintain the RCR training *status quo* and as such, prevent a wider and deeper level of engagement in ethical reasoning and its purposeful cultivation across the scientists' career, and irrespective of how direct the investigator's contact is with human or animal research subjects, across scientific disciplines. Although this chapter is focused on the weaknesses of the NIH RCR training paradigm for BBD training and practice, it is possible that other institutions have similar general ethics training programs, equally deeply entrenched and also neither relevant nor effective.

4 An Alternative RCR Training Paradigm

Lessons from the Dominant RCR Training Paradigm: While we fully agree with the NIH Basic Principles, and with the importance ascribed to the topical areas outlined by the NIH as important for responsible conduct in research (see also National Academy of Sciences and National Academy of Engineering, and Institute of Medicine 2009), the foregoing suggests that the NIH RCR training paradigm

and its implementation (and particularly its exceptions), have no chance to support the ideal that "... (t)he entire community of scientists and engineers benefits from diverse, ongoing options to engage in conversations about the ethical dimensions of research and (practice)" (Kalichman 2013: 13). Nor is there any possibility of the "... integration and institutionalization of established mechanisms of reflection ... around the process of research and innovation." (Owen et al. 2012: p. 755). Instead, the NIH RCR training paradigm specifically articulates that only some of those being trained with public funds should receive training in responsible conduct of research – thereby implying that others, including those in BBD, should not. Without a majority of researchers receiving, and/or competently providing, this training, only a minority of the scientific community would be able to engage, much less benefit, from ongoing conversations about the ethical dimensions of their research and practice. The one-time "inoculation" that RCR training often seems designed to provide (see Novossiolova and Sture 2012) is not consistent with the NIH Basic Principles and is also not capable of supporting even the initiation of, much less the ongoing options to engage in, "conversations about the ethical dimensions of research and (practice)" (Kalichman 2013). The dominance of this paradigm also supports a culture where such conversations are, and are made to seem, unnecessary and unimportant. Therefore, an alternative training paradigm is required in order to promote a culture of ethical practice in and with BBD. We feel strongly that promoting training to support ethical professional practice in BBD through the university and biomedical research infrastructure around the world will create the culture of ethical BBD practice that will dictate the norms for BBD practice *outside* of academia.

Within academia, *not all institutions, programs, and faculty follow the "NIH RCR training" paradigm*, but where this paradigm dictates RCR training, ongoing conversations about ethics in science – whether biomedical or not, and involving big *or small* data – are not supported (see also Novossiolova and Sture 2012). Quite apart from its internal inconsistency, the dominant NIH RCR training paradigm is not sufficient to introduce, promote, *and sustain* research integrity throughout a scientific career (Tractenberg and FitzGerald 2012, 2015; Tractenberg et al. 2014; Tractenberg et al. 2015). As an alternative to the NIH RCR training paradigm, we articulated a developmental paradigm comprised of learnable, improvable skills that can be deployed around any topics – including those that have not yet been encountered – and whose development and growth can be documented over time. These learnable, improvable skills are required for ethical reasoning – which is the foundation of our paradigm. Derived from public scholarship on ethical decision-making, the knowledge, skills and abilities (KSAs) in our formulation (and teaching) of ethical reasoning are: (identification and assessment of one's) prerequisite knowledge; recognition of a moral issue; identification of relevant decision-making frameworks; identification and evaluation of alternative actions; making & justifying a decision (about the moral issue); and reflection on the decision (http://www.scu.edu/ethics/; see also Kligyte et al. 2008; Hollander and Arenberg 2009 for similar constituents of ethical reasoning derived from non-ethics perspectives). Our work –and this paradigm – is based on two ideas:

1. Ethics education should *inculcate* – seed and support the development of – a professional and ethical identity that can then grow over a *career* in science or practice (or both); and
2. Increasing faculty and student awareness of alternative RCR training paradigms that are more effective, even if they are also more demanding, than just regulatory compliance via one-time "RCR training", can motivate these individuals to become agents of culture change who promote these alternatives.

These two ideas represent the lessons that we have learned as we developed an alternative to the NIH RCR training paradigm (initiated in Tractenberg and FitzGerald 2012). Our recent research and teaching around ethics education is specifically focused on the initiation of a developmental process in ethical reasoning that can span an entire scientific career, challenging this dominant paradigm on multiple levels, starting with the concept that one course every 4 years is sufficient to both initiate and sustain ethical awareness over the course of a career. However we also challenge the idea that only some scientists must have training in ethical reasoning. Instead, we articulate that training in ethical reasoning is relevant in any discipline because it can inculcate future practitioners with the sense of professional identity that can motivate and sustain interest in cultivating the habits of mind that experts in that discipline/profession possess. This paradigm does not limit "training in responsible conduct" to any group, or to any type of data or research – or even to future practitioners of "science" or "research". It does permit an emphasis on the different decision-making frameworks and sources of moral issues that can be interesting and important within those disciplines that comprise more philosophy, but does not require that emphasis.

We have shown that a semester course can be developed to initiate ethical reasoning for students in computing or in statistics (Tractenberg et al. 2015); both of these courses, while providing training in ethical reasoning that can support future growth in these skills (Tractenberg et al. in review) and initiating familiarity with the codes of conduct (reviewed below) for these disciplines, also meet (and probably exceed) the NIH requirements for training in RCR. This is true relative to both the list of NIH topics *and* the NIH Basic Principles. Moreover, our recent work suggests that a single course in ethical reasoning with just the NIH RCR topics list (i.e., *without* a discipline-specific code of conduct) is supportive of the NIH Basic Principles, while both introducing new and developing scientists to, and also promoting *sustainability* of, research integrity and a sense of responsibility in their conduct of research. That is, our preliminary results (based on three successive cohorts of 3–5 PhD students and PhD-holding auditors; reported in Tractenberg et al. in review) suggest that our alternative RCR training paradigm addresses the three weaknesses in the NIH RCR training paradigm:

1. All scientists can receive –and engage in- discipline-relevant training in the responsible conduct of research.
2. There are achievable requirements to document the capacity to reliably train others in RCR or research ethics, and after one semester course and a further

few months of reflection and portfolio development, our students can document their growth towards this objective.
3. Our preliminary evidence is that our RCR training paradigm leads to both the growth in knowledge around the NIH topics lists that meets the NIH requirement, and also leads to the capacity to continue building on that knowledge beyond the end of the course and in other aspects of their professional lives.

Because they may not be as invested in, experienced with, or driven by the topics-specific one-course-done paradigm for "training in RCR", quantitatively-trained participants may not have many opportunities to learn about, and determine how to prioritize or make decisions around the ethical, legal, and social implications (ELSI) of working with biomedical big data. Furthermore, while the Basic Principles of the NIH RCR paradigm (namely that training in ethical conduct should be required for all investigators (NIH 2009)) *should* be adopted in and for Big Data, the actual RCR training *paradigm* itself does not work and should be avoided. Instead, individuals coming to biomedical big data from computing and statistics have established, articulated ethical guidelines for professional practice; our ethical reasoning training paradigm can leverage these codes of conduct to ensure that *all* participants can engage in training that supports professional identity development, promotes ethical reasoning that can be sustained beyond the end of the training and into areas/experiences that were not originally taught about, and is provided by individuals whose abilities to train in ethical reasoning are documented. Two semester courses initiating this training are presented and described in Tractenberg et al. (2015); the integration of a code of conduct into a course, a program of courses, or a curriculum are outlined in Tractenberg (Tractenberg 2016b). Background on several existing professional guidelines for ethical practice is outlined below, to support the argument that they comprise similar constructs to the NIH RCR topics list –but are discipline specific and relevant for the initiation and development of professional identity. Non-members of these Associations can obtain the Ethical Guidelines, and there are many resources openly and freely available for creating training opportunities outside of university settings. University-affiliated members of these associations might also be engaged to create relevant training to support ethical reasoning with the disciplinary Ethical Guidelines that are most relevant for any non-university application in BBD. The NIH RCR training paradigm should be avoided, but there are many resources that can be leveraged outside of academia in order to promote a culture of ethical practice with BBD.

5 Ethical Principles in Statistical Analysis/Research

As was outlined in Tractenberg et al. (2015), those who are being trained as Big (or small) Data scientists might be completely unaware that there exist clear and concrete codes of conduct for professional practice in these disciplines; this includes those whose instruction and experience is likely to emphasize the analysis of data

(e.g., programs aligned with the American Statistical Association (ASA), Royal Statistical Society (RSS), or International Statistics Institute (ISI)) or computing, managing, and information science/technological aspects of data (e.g., programs aligned with the Association for Computing Machinery (ACM) or Institute of Electrical and Electronics Engineers (IEEE)). This could be because these codes of conduct are not currently taught in any university curriculum; e.g., Lee et al. (2015) estimated that 35 % of all universities in the US with undergraduate or graduate programs in statistics or biostatistics require "*some* ethics training for at least *some* students" (emphasis added), but there is no indication that any of this ethics training in the US includes or mentions the ASA Ethical Guidelines for Statistical Practice (the current (2015) Director of the ASA, Chair of the ASA Professional Ethics Committee, and the Chair of the working group that revised the ASA Ethical Guidelines for Statistical Practice in 1999 have each confirmed that there are no formal training opportunities involving the ASA Guidelines). However, it might also be the case that, because professional society guidelines are discipline (and professional society) –specific, students and faculty who are not specifically or tightly aligned with, or members of, these Associations might not identify with their Guidelines or recognize their utility or even their existence.

The alignment of the Beauchamp and Childress (1983) principles, which underpin many or most of the frameworks within which a majority of biomedical researchers in the United States are trained in the responsible conduct of research, with the ethical principles that have been outlined for those whose engagement in biomedical research may focus on, or be primarily driven by, big data, is superficial at best. Tractenberg (2013) demonstrated that, although represented as supportive of ethics training for quantitative researchers by Rosnow and Rosenthal (2011) (in their Table 1), the Beauchamp and Childress (1983) principles are not consistent with the five dimensions of data analytic and reporting standards (transparency; informativeness; precision; accuracy; and theoretical groundedness) with which they are grouped in that 2011 chapter. Given their total orthogonality in origin and purpose, this is not surprising. However, in the Appendix Table A1, Tractenberg (2013) outlines how the current NIH RCR topical areas (representing, but not specific to, the Beauchamp & Childress principles) can very clearly be mapped *together* with the eight main domains of the ASA Ethical Guidelines for Statistical Practice (professionalism; responsibilities to funders, clients & employers assuring that statistical work is suitable; responsibilities in publications and testimony; responsibilities to research subjects; responsibilities to research team colleagues; responsibilities to other statisticians or statistical practitioners; responsibilities regarding allegations of misconduct; and responsibilities of employers; described in further detail below).

The American Statistical Association approved the 2016 revised version of the Ethical Guidelines for Statistical Practice (the author chaired the revision task force, served as Vice Chair of this ASA Committee and takes over as chair July 2016). The content of each of the general areas was revised for clarity and updating, and the eight topic areas (ASA 2016) are given here, with each Principle comprising

4–11 focal elements articulated in the full Code online (http://community.amstat.org/ethics/home):

A. **Professional Integrity & Accountability** points out the need for competence, judgment, diligence, self-respect, and worthiness of the respect of other people.
B. **Integrity of data and methods (formerly "Responsibilities in Publications and Testimony")** addresses the need to report sufficient information to give readers, including other practitioners, a clear understanding of the intent of the work, how and by whom it was performed, and any limitations on its validity.
C. **Responsibilities to Science/Public/Funder/Client** discusses the practitioner's responsibility for assuring that statistical work represents "good science", suitable to the needs and resources of those who are paying for it, that funders understand the capabilities and limitations of statistics in addressing their problem, and that the funder's confidential information is protected.
D. **Responsibilities to Research Subjects** describes requirements for protecting the interests of human and animal subjects of research-not only during data collection but also in the analysis, interpretation, and publication of the resulting findings.
E. **Responsibilities to Research Team Colleagues** addresses the mutual responsibilities of professionals participating in multidisciplinary research teams.
F. **Responsibilities to Other Statisticians or Statistics Practitioners** notes the interdependence of professionals doing similar work, whether in the same or different organizations. Basically, they must contribute to the strength of their professions overall by sharing non-proprietary data and methods, participating in peer review, and respecting differing professional opinions.
G. **Responsibilities Regarding Allegations of Misconduct** addresses the sometimes painful process of investigating potential ethical violations and treating those involved with both justice and respect.
H. **Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners** encourages employers and clients to recognize the highly interdependent nature of statistical ethics and statistical validity. Employers and clients must not pressure practitioners to produce a particular "result," regardless of its statistical validity. They must avoid the potential social harm that can result from the dissemination of false or misleading statistical work.

The Royal Statistical Society (RSS) first published its Code of Conduct in 1993, and it was revised in 2014 (RSS 2014) The Code is recommended for all RSS members (called fellows) and is "mandatory for Chartered Statisticians" (CStat), the designation from the RSS that requires at least 5 years of professional experience, along with other application materials; and also for students who seek and are awarded the GradStat designation, which identifies graduate student members who are committed to the Code of Conduct and also to a formal continuing professional development policy. The RSS domains and the constituent rules are shown here; each rule has a list of specific principles articulated in the full Code online:

The Public Interest

1. Fellows should always be aware of their overriding responsibility to the public good; including public health, safety and environment.
2. Fellows shall in their professional practice have regard to basic human rights and shall avoid any actions that adversely affect such rights.

Obligation to Employers and Clients

3. Fellows shall carry out work with due care and diligence in accordance with the requirements of the employer or client.
4. Fellows shall respect any agreements of confidentiality entered into with an employer or client.
5. Fellows should not allow their name to be attributed to work that they have either not contributed to or which presents their contribution in a misleading way.

Obligation to the Profession and the Society

6. Fellows shall uphold the reputation of the Profession and the Society
7. Fellows shall seek to advance knowledge and understanding of statistical science and advocate its use.
8. Fellows shall act with integrity towards fellow statisticians and to members of other professions with whom they collaborate.
9. Fellows shall take personal responsibility for work bearing their name.

Professional Competence and Integrity

10. Fellows shall strive to act with honesty and integrity in all aspects of their professional life.
11. Fellows shall undertake continuing professional development (CPD) in accordance with the CPD Policy of the Society in order to maintain or upgrade their professional knowledge and skill and maintain awareness of technical developments, procedures and standards which are relevant to their field, and shall encourage others to do likewise.
12. Fellows shall seek to conform to recognised good practice including quality standards which are in their judgment relevant, and shall encourage others to do likewise.
13. Fellows shall report to the Society any criminal convictions against them in respect of violence, dishonesty or professional misconduct; or upon becoming bankrupt or disqualified as Company Director.

The International Statistics Institute (ISI) approved its most recent version of the Declaration, originally adopted in 1985, on Professional Ethics in 2010 (http://www.isi-web.org/images/about/Declaration-EN2010.pdf). In its preamble, the Declaration states that "The aim of this declaration is to enable the statistician's individual ethical judgments and decisions to be informed by shared values and experience, rather than by rigid rules imposed by the profession." (p. 3) There are 12 Ethical Principles outlined:

1. Pursuing Objectivity
2. Clarifying Obligations and Roles
3. Assessing Alternatives Impartially Available methods and procedures should be considered and an impartial assessment provided to the employer, client, or funder of the respective merits and limitations of alternatives, along with the proposed method.
4. Conflicting Interests
5. Avoiding Preempted Outcomes
6. Guarding Privileged Information
7. Exhibiting Professional Competence
8. Maintaining Confidence in Statistics
9. Exposing and Reviewing Methods and Findings
10. Communicating Ethical Principles
11. Bearing Responsibility for the Integrity of the Discipline
12. Protecting the Interests of Subjects

6 Ethical Principles in Computing

Like other professional associations, the Association of Computing Machinery (ACM) developed its Code of Ethics and Professional conduct to represent a commitment to "ethical professional conduct" by all ACM members, and this explicitly includes students, associates, and voting members (ACM 1992). There are General Moral Imperatives, formulated according to the statement, "As an ACM member I will . . . "

1.1 Contribute to society and human well-being.
1.2 Avoid harm to others.
1.3 Be honest and trustworthy.
1.4 Be fair and take action not to discriminate.
1.5 Honor property rights including copyrights and patent.
1.6 Give proper credit for intellectual property.
1.7 Respect the privacy of others.
1.8 Honor confidentiality.

There are two more sections (see below) and the fourth section contains two elements, both also following the statement, "As an ACM member I will . . . ":

4.1 Uphold and promote the principles of this Code.
4.2 Treat violations of this code as inconsistent with membership in the ACM.

The two middle sections are "More Specific Responsibilities" (Section 2) and "Organizational Leadership Imperatives" (Section 3), which are also formulated according to statements more specific to professional practice: "As an ACM computing professional I will . . . " and "As an ACM member and an organizational leader, I will . . . ", respectively. It is articulated that "leader" includes individuals with educational responsibilities.

2.1 Strive to achieve the highest quality, effectiveness and dignity in both the process and products of professional work.
2.2 Acquire and maintain professional competence.
2.3 Know and respect existing laws pertaining to professional work.
2.4 Accept and provide appropriate professional review.
2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.
2.6 Honor contracts, agreements, and assigned responsibilities.
2.7 Improve public understanding of computing and its consequences.
2.8 Access computing and communication resources only when authorized to do so.
3.1 Articulate social responsibilities of members of an organizational unit and encourage full acceptance of those responsibilities.
3.2 Manage personnel and resources to design and build information systems that enhance the quality of working life.
3.3 Acknowledge and support proper and authorized uses of an organization's computing and communication resources.
3.4 Ensure that users and those who will be affected by a system have their needs clearly articulated during the assessment and design of requirements; later the system must be validated to meet requirements.
3.5 Articulate and support policies that protect the dignity of users and others affected by a computing system.
3.6 Create opportunities for members of the organization to learn the principles and limitations of computer systems.

These codes all articulate practice-specific principles for professional activities, and – especially the ACM code – vary in terms of the relevance and level of sophistication that any given practitioner would be expected to exhibit with respect to each principle given that individual's career stage and level of responsibilities. The codes –which are only excerpted here – all contain preambles that outline the intention of the code and its applicability for the practitioner, and although each is of great importance to the Association, *none* is currently used as a basis for professional identity development within the field. Moreover, if trainees within these domains are federally funded, and if their institution deems the NIH RCR training paradigm to be required, they are more likely to have the sorts of training opportunities shown in Table 1 than to have any training opportunity that is as specific to their professional practice, or to their discipline, as these codes are. This is yet another reason for arguing that the NIH RCR training paradigm should be *avoided* and that new training paradigms to promote professionalism and ethical practice for BBD should be developed that emphasize these practical, practice-specific codes. The ethical reasoning training paradigm that we have outlined is (of course!) recommended, but we recommend and would welcome any training paradigm that promotes universal and ongoing engagement with the codes/guidelines of the profession, initiates sustainable training (i.e., training that continues beyond the end of the course and can be applied outside the specific topics

of the course), and requires some level of documentation of competence to instruct. Departments or programs that contemplate the integration of a developmental trajectory relating to their (or any relevant) professional code of conduct can begin to support Kalichman's (2013) ideal ("... (t)he entire community of scientists and engineers benefits from diverse, ongoing options to engage in conversations about the ethical dimensions of research and (practice),"). They would also instantiate Shulman's (2008) description, to "*reflect the character of inquiry, the nature of community, and the ways in which research and teaching are conducted in that particular discipline or disciplinary intersection*" – thereby creating and supporting a culture of professional ethics for themselves and their students. If achieved across students preparing to engage with BBD, this type of training paradigm would also support the evolution of "... integration and institutionalization of established mechanisms of reflection ... around the process of research and innovation." (Owen et al. 2012: p. 755). Thus, the NIH RCR training paradigm may be convenient, but it is not effective nor does it support any of the considerations in RRI.

7 Support for Ongoing Reflection on Professional Obligations

As mentioned, Tractenberg and FitzGerald (2012) described ethical reasoning as a learnable, improvable set of knowledge, skills and abilities (KSAs) that can be learned and practiced, and applied to many different ethical challenges and decisions (including those that didn't exist or were unknown to the reasoner when the KSAs were initially learned and practiced). Our approach to ethics education for biomedical researchers was developed by a cognitive scientist (the author, RET) working with an ethicist (KT FitzGerald). It differs from the NIH RCR training paradigm on many levels, one being that it does represent considerably more work for the learner *and instructor* than anything shown in Table 1. Specifically, this approach relies on formative assessment that is so important to ethics education- and that can be so difficult (e.g., Keefer and Davis 2012) for many faculty in (and outside of!) Biomedical Big Data to provide. The purpose of the semester course in ethical reasoning that we developed is for more expert ethical reasoners (instructors) to guide the development and practice of the KSAs by less-expert ethical reasoners until learners have all reached a common level of awareness about their reasoning (see Tractenberg et al. 2015). If students wish to, they can continue to develop these reasoning KSAs –either seeking or creating opportunities for continuing to grow and develop these particular skills throughout their careers. If *these* students eventually engaged in teaching within institutional RCR training opportunities such as those shown in Table 1, they *would* be able to demonstrate and document their increasing sophistication of these KSAs – presumably by immediately restructuring the RCR training opportunities in which they engage to be relevant and supportive of professionalism. Through our course, students learn to attend to, and reflect on, their

own skills, i.e., become metacognitive around their reasoning (Tractenberg et al. in review). In addition to being constructive and authentic, features that promote adult learning (Knowles and Holton 2005), our paradigm also represents the elements most supportive of "effective" ethics education (May and Luth 2013; see also Antes et al. 2009) – not just factual "ethics" knowledge but also, reasoning in the face of uncertainty (see Mumford et al. 2008) and a sense of purpose for engaging in self-directed learning (see Paterson et al. 2002; Ambrose et al. 2010, pp. 190–216). The argument in this chapter, and articulated by Tractenberg et al. (2015), is that an appreciation for the role of 'expert habits of mind' that derive *from the discipline*, i.e., from discipline-specific codes for ethical professional practice, can engage learners in both the immediate learning enterprise (e.g., the course) and its ongoing pursuit. Although our evidence suggests that our ethical reasoning training paradigm overcomes the three main objections to the NIH RCR training paradigm, it is definitely more challenging to implement than that dominant paradigm; in our work to study, document, and understand the strengths and limitations of our paradigm, we have encountered several barriers to its implementation, outlined in the next section.

8 Barriers to Rejecting the NIH RCR Training Paradigm

The author has contacted graduate (PhD and MS) program directors and department chairs around the United States, and one American and one international data centre, none of which has any RCR training requirements (neither from the institution nor from their major funders). None of these program directors or department chairs was in a situation where the NIH RCR training paradigm is actually dominant – because there is no requirement currently for any training in ethics or preparation for professional practice beyond the degree-completion requirements of the curriculum. While all of these programs would label themselves as firmly within "scientific" disciplines, none of them actively promotes the notion of the "responsibilities, including those associated with the concepts of research integrity", that Owen et al. (2012) describe as foundational for RRI – upon which "new responsibilities" are added in the new era of RRI. Instead, this sample of programs in statistics, biostatistics, economics, physics, and computer science were unwilling to engage their students in training in ethics or professional ethical guidelines, because of one of the following reasons:

1. This program/discipline does not need training in ethics (N = 2/10).
2. I have never thought about/heard of training in ethics (either at all or for my field) (2/10).
3. I am intrigued, but I don't think it could work (to offer a whole course just on ethics; to train faculty to provide this training; to ask faculty to take the time to learn the paradigm so they can teach discipline-specific ethics) (4/10; 2/4 said they would like to learn more).

4. If I could make time for a course in discipline-specific ethics, I would actually dedicate that time to teaching a new method – our students will get training in context-specific ethics when they get jobs with their new degrees (2/10).

By contrast, when we approached colleges of Engineering (N = 3) around the region (mid-Atlantic US) to inquire about their interest in trying our new program, we received strong support and interest in a new way to both integrate training in professionalism, professional identity, and professional ethics into the curricula. All Engineering faculty we spoke to stated that their programs undergo accreditation procedures and ethics is a key component of these procedures – and most also stated that their faculty would be interested in learning new methods for promoting a sense of professional ethics across the curriculum. The lesson here is that, not only is the NIH RCR training paradigm difficult to dislodge, but also that efforts to promote other, new, or different ethics training paradigms must be *plausible* for the department and program leadership, and its integration must be deemed important for student (graduate/alumnus) success. Because the current system "works" in the sense that no training is required or compliance can be documented without additional resources (time, effort, courses, etc.), there is a great deal of inertia to overcome. These disciplines are, as noted earlier, making critical contributions to research and innovation involving BBD; their professional preparation should therefore be required to include some formal instruction around their responsibilities to begin to learn how to reflect on the ELSI of their work and contributions.

9 Discussion

As outlined above, computing machinery and statistics have clearly-articulated ethical guidelines/guidelines for professional practice- as do many other domains. These are not specific to the scientists' responsibilities relating to "research integrity" – although that is subsumed – but do represent a substantial, authentic, mode for introducing training in responsible conduct within their scientific and technical domains. The NIH RCR training paradigm specifically identifies these codes as insufficient to support training in the responsible conduct of biomedical research, although even the superficial evaluation of the content presented here – which is greatly abridged- shows that this is *not* the case (see also Tractenberg 2013; and the syllabi presented in Tractenberg et al. 2015). Because these professional associations each assert that all members of the target profession have an obligation to know and follow the code or guidelines, their potential to become more widespread and more consistently delivered is *greater* than that of the NIH RCR training paradigm, particularly since NIH does not require any RCR training for those whose work does not involve "human subjects". We have argued elsewhere how and that the professional code of conduct for an association can be leveraged to ensure that quantitatively-trained participants in work with biomedical big data know and understand –or are able to reason around – the ethical, legal, and

social implications (ELSI) of their involvement in this research (e.g., Tractenberg 2013, 2016a; Tractenberg and FitzGerald 2015). Promoting a sense of professional identity can do much – far more than the NIH RCR training paradigm – to promote a culture of ethical practice in BBD. With this culture would come a broader base of support for the "... integration and institutionalization of established mechanisms of reflection ... around the process of research and innovation" envisioned by Owen et al. (2012), and Kalichman's (2013) ideal that "... (t)he entire community of scientists and engineers benefits from diverse, ongoing options to engage in conversations about the ethical dimensions of research and (practice)". The NIH RCR training paradigm has no hope of promoting such widespread engagement with either the reflection or the conversation that these ideals embody.

This chapter has outlined how "ethics training" that emphasizes respect for persons; beneficence; and justice; or privacy and informed consent, will fail to prepare quantitatively-trained practitioners to competently address the ELSI that are of primary concern for those engaging in Big Data analysis and management. Moreover, Biomedical Big Data should ignore the model represented by the NIH RCR training paradigm because of its internal inconsistency; focus on users *outside* of Biomedical Big Data applications; emphasis on completion, rather than engagement; and lack of value or support for, or promotion of, ongoing and deepening responsibility in the conduct of research. Instead, ethics training should embrace and embody the promotion of professional identity formation that teaching, and exploring the applications of, ethical guidelines for professional practice and conduct can provide to those who are training to join a profession like statistics or computer science. The introduction of all potential practitioners in Biomedical Big (and small) Data to the habits of mind that experts in those fields exhibit can be integrated into (i.e., throughout) the degree programs that seek or claim to prepare trainees for these professions and research with and in Biomedical Big Data (examples from the ASA outlined in Tractenberg 2016b).

Formal and pervasive integration of these guidelines into the preparation for practice with Big Data in a biomedical context is feasible; both statistics and computer science associations argue for the inclusion of training in ethics for undergraduates participating in degree programs that are aligned with preparing future professionals in these disciplines (see Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and IEEE Computer Society (2013); and Horton and the ASA undergraduate curriculum workgroup (2014). Tractenberg (2016b) outlines how the ASA Ethical Guidelines for Statistical Practice can be integrated into existing courses, programs or sequences of course, and curricula; Tractenberg et al. (2015) published semester course syllabi that teach either the ACM General Moral Imperatives or the ASA Ethical Guidelines topical areas. This integration can help future investigators deal with challenges that have not yet arisen, because it can support ongoing reflection on professional obligations that can be deployed in a wide range of ELSI -including those not yet identified.

> *Both scholarship and teaching in any field reflect the character of inquiry, the nature of community, and the ways in which research and teaching are conducted in that particular discipline or disciplinary intersection*. (Shulman 2008, p. xii).

Because of the team-based nature of modern biomedical research, it is important to promote the conceptualization of "Biomedical Big Data" as a disciplinary intersection. Specifically, the ethical guidelines for professional practice of any of the disciplines at that intersection – and especially not solely those of the NIH RCR training paradigm – should dictate the level of familiarity with all relevant disciplinary obligations. Preparing scientists to engage competently in ongoing conversations around ethical issues in biomedical Big Data requires purposeful, discipline-relevant, and developmental training that can come from, and support, a culture of ethical biomedical research and practice with Big Data.

Acknowledgment The Author was supported by a grant (Award 1237590) from the National Science Foundation.

References

Ambrose, S.A., M.W. Bridges, M. DiPietro, M.C. Lovett, and M.K. Norman. 2010. *How learning works: Seven research-based principles for smart teaching*. San Francisco: Wiley.

American Statistical Association. 2016. *Revised ethical guidelines for statistical practice*. Downloaded from http://community.amstat.org/ethics/home. Accessed 3 May 2016.

Antes, A.L., S.T. Murphy, E.P. Waples, M.D. Mumford, R.P. Brown, S. Connelly, and L.D. Devenport. 2009. A meta-analysis of ethics instruction effectiveness in the sciences. *Ethics and Behavior* 19(5): 379–402.

Antes, A.L., X. Wang, M.D. Mumford, R.P. Brown, S. Connelly, and L.D. Devenport. 2010. Evaluating the effects that existing instruction on responsible conduct of research has on ethical decision making. *Academic Medicine* 85: 519–526.

Association of Computing Machinery (ACM). 1992. *ACM code of ethics and professional conduct*. Downloaded from http://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct#CONTENTS.

Beauchamp, T.L., and J.F. Childress. 1983. *Principles of biomedical ethics*, 2nd ed. Oxford: Oxford University Press.

Boyd, D., and K. Crawford. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society* 15(5): 662–679.

Department of Health and Human Subjects. 2009. *Code of Federal Regulations, 45 CFR 46*. Downloaded from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html. 6 June 2015.

Elias, P. 2014. *A European perspective on research and big data analysis. Privacy, big data, and the public good: Frameworks for engagement*. Cambridge, UK: Cambridge University Press. http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/privacy-big-data-and-public-good-frameworks-engagement.

Fanelli, D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 4(5), e5738. doi:10.1371/journal.pone.0005738.

Fang, F.C., R.G. Steen, and A. Casadevall. 2012. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences (PNAS)* 109(42): 17028–17033. Downloaded from http://www.pnas.org/content/109/42/17028.full.pdf. Accessed on 6 June 2015.

Hollander, R., and C.R. Arenberg (eds.). 2009. *Ethics education and scientific and engineering research*. Washington: National Academy of Engineering.

Horton, N., and the American Statistical Association Undergraduate Guidelines Workgroup. 2014. Curriculum guidelines for undergraduate programs in statistical science. Downloaded from http://www.amstat.org/education/pdfs/guidelines2014-11-15.pdf. Accessed 14 Feb 2015.

International Statistics Institute. 2010. Declaration on professional ethics. Downloaded from http://www.isi-web.org/images/about/Declaration-EN2010.pdf. Accessed 12 May 2015.

Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) & IEEE Computer Society. 2013. Computer science curricula 2013. Downloaded from https://www.acm.org/education/CS2013-final-report.pdf. Accessed 21 Nov 2014.

Kalichman, M. 2013. Why teach research ethics? In *Practical guidance on science and engineering ethics education for instructors and administrators*, ed. National Academy of Engineering, 5–16. Washington: National Academies Press.

Kalichman, M.W., and D.K. Plemmons. 2007. Reported goals for responsible conduct of research courses. *Academic Medicine* 82: 846–852.

Keefer, M., and M. Davis. 2012. Curricular design, instruction, and assessment in professional ethics education: Some practical advice. *Teaching Ethics* 12: 81–90.

Kligyte, V., R.T. Marcy, S.T. Sevier, E.S. Godfrey, and M.D. Mumford. 2008. A qualitative approach to Responsible Conduct of Research (RCR) training development: Identification of metacognitive strategies. *Science and Engineering Ethics* 14(1): 3–31.

Knowles, M.S., and E.F. Holton III. 2005. *The adult learner*, 6th ed. Burlington: Elsevier.

Lee, L.M., F.A. McCarty, and T.R. Zhang. 2015. Ethical numbers: Training in US graduate statistics programs, 2013–2014. *The American Statistician* 69(1): 11–16. doi:10.1080/00031305.2014.997891.

May, D.R., and M.T. Luth. 2013. The effectiveness of ethics education: A quasi-experimental field study. *Science and Engineering Ethics* 19(2): 545–568.

Mumford, M.D., S. Connelly, R.P. Brown, S.T. Murphy, J.H. Hill, A.L. Antes, E.P. Waples, and L.D. Devenport. 2008. A sensemaking approach to ethics training for scientists: Effects on ethical decision-making. *Ethics and Behavior* 18: 315–339.

Mumford, M.D., S. Connelly, S.T. Murphy, L.D. Devenport, A.L. Antes, R.P. Brown, J.H. Hill, and E.P. Waples. 2009. Field and experience influences on ethical decision-making in the sciences. *Ethics and Behavior* 19(4): 263–289.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2009. *On being a scientist: A guide to responsible conduct in research*, 3rd ed. Washington: National Academies Press.

National Institutes of Health. 1989. NIH Guide 18(45 December). Vol. 18(45). December 22. Downloaded from http://grants.nih.gov/grants/guide/historical/1989_12_22_Vol_18_No_45.pdf. Accessed 6 June 2015.

National Institutes of Health. 2009. Update on the requirement for instruction in the responsible conduct of research. NOT-OD-10-019. http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-10-019.html. Downloaded from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-10-019.html. Accessed 29 Nov 2009.

Novossiolova, T., and J. Sture. 2012. Towards the responsible conduct of scientific research: is ethics education enough? *Medicine, Conflict, and Survival* 28(1): 73–84.

Owen, R., P. Macnaghten, and J. Stilgoe. 2012. Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy* 39(6): 751–760. doi:10.1093/scipol/scs093.

Paterson, M., J. Higgs, S. Wilcox, and M. Villeneuve. 2002. Clinical reasoning and self-directed learning: Key dimensions in professional education and professional socialisation. *Focus on Health Professional Education: A Multi-Disciplinary Journal* 4(2): 5–21.

Rosnow, R.L., and R. Rosenthal. 2011. Ethical principles in data analysis: An overview. In *Handbook of ethics in quantitative methodology*, ed. A.T. Panter and S.K. Sterba, 37–59. New York: Taylor and Francis.

Royal Statistical Society. 2014. Code of conduct. Downloaded from http://www.rss.org.uk/RSS/Join_the_RSS/Code_of_conduct/RSS/Join_the_RSS/Code_of_conduct.aspx?hkey=3170e215-12c6-4948-b023-e7253a4600a8. Accessed 13 May 2015.

Schmaling, K.B., and A.W. Blume. 2009. Ethics instruction increases graduate students' responsible conduct of research knowledge but not moral reasoning. *Accountability in Research* 16: 268–283.

Shulman, L.S. 2008. Foreward. In *The formation of scholars: Rethinking doctoral education for the twenty first century*, ed. G.E. Walker, C.M. Golde, L. Jones, A.C. Bueschel, and P. Hutchings, ix–xiii. San Francisco: Jossey-Bass.

Steinmann, M., J. Shuster, J. Collmann, S. Matei, R.E. Tractenberg, K. FitzGerald, G. Morgan, and D. Richardson. 2015. Embedding privacy and ethical values in Big Data technology. In *Transparency on social media – Tools, methods and algorithms for mediating online interactions*, ed. S.A. Matei, M. Russell, and E. Bertino, 277–301. New York: Springer.

Tractenberg, R.E. 2013. Ethical reasoning for quantitative scientists: A mastery rubric for developmental trajectories, professional identity, and portfolios that document both. Proceedings of the 2013 Joint Statistical Meetings, Montreal.

Tractenberg, R.E. 2016a. Integrating ethical reasoning into preparation for participation to work in/with Big Data through the Stewardship model. In *Ethical reasoning in big data: An exploratory analysis*, ed. J. Collmann and S. Matei, 185–192. New York: Springer.

Tractenberg, R.E. 2016b. Institutionalizing ethical reasoning: Integrating the ASA's Ethical Guidelines for Professional Practice into course, program, and curriculum. In *Ethical reasoning in big data: An exploratory analysis*, ed. J. Collmann and S. Matei, 115–139. New York: Springer.

Tractenberg, R.E., and K.T. FitzGerald. 2012. A Mastery Rubric for the design and evaluation of an institutional curriculum in the responsible conduct of research. *Assessment and Evaluation in Higher Education* 37(7–8): 1003–1021.

Tractenberg, R.E., and K.T. FitzGerald. 2015. Responsibility in the conduct of quantitative sciences: Preparing future practitioners and certifying professionals. Presented at the 2014 Joint Statistical Meetings, Boston; to appear in Proceedings of the 2015 Joint Statistical Meetings, Seattle.

Tractenberg, R.E., Russell, A., Morgan, G., et al. 2015. Amplifying the reach and resonance of ethical codes of conduct through ethical reasoning: Preparation of Big Data users for professional practice. *Science and Engineering Ethics*. http://link.springer.com/article/10.1007%2Fs11948-014-9613-1

Tractenberg, R.E., K.T. FitzGerald, J. Collmann, and J. Giordano. 2014. Big data impact upon neuroS/T use for influence and deterrence. In: *Strategic Multilayer Assessment Group whitepaper on leveraging neuroscientific and neurotechnological (NeuroS&T) developments with focus on influence and deterrence in a networked world*, eds. D. DiEuliis, W. Casebeer, J. Giordano, N. Wright, and H. Cabayan, Joint Staff, J3, DDGO, OSD/ASD (R&E)/RSD/RRTO [Governmental White Paper].

Tractenberg, R.E., FitzGerald, K.T., and Collmann, J. in review. *Evidence of sustainable learning with the Mastery Rubric for Ethical Reasoning*.

Part VI
Foresight

The Ethics and Politics of Infrastructures: Creating the Conditions of Possibility for Big Data in Medicine

Linda F. Hogle

Abstract The vision of creating a more comprehensive understanding of human health and disease calls for collecting ever-greater volumes of information about individuals, in continuous, real-time streams, and from sources outside of the clinic. Such data is heterogenous and high-dimensional, requiring the use of big data analytics. Big data has been granted considerable perceived authority to solve problems in healthcare and biomedicine. At the same time, there is potential for tremendous impact on social and political life more broadly. For this reason, it is important to elucidate less-visible ethical issues related to the infrastructures being built to support big data projects in biomedical science and clinical medicine. The constellation of changes in laws, institutional arrangements, and new forms of expertise being brought together are reordering relations among patients, clinical and family caregivers, researchers and payers, with potentially long-term effects. For example, conventional concepts of autonomy are challenged when data is collected ubiquitously and passively, and notions of expertise are provoked when 'non-medical' experts (including patients themselves) participate more directly in processes of defining health, illness, and care. In the process, the distinction between research and clinical activities (which have been conceptually kept apart for decades) becomes blurred, and the definition of 'research subject' is confounded.

1 Introduction

> "... in the not-too-distant future, each patient will be surrounded by a 'virtual cloud' of billions of data points that will uniquely define their past medical history and current health status. Furthermore, it will be possible to mine the billions of data points from hundreds of millions of individuals to generate algorithms to help predict the future clinical needs for each patient." (Hood and Friend 2011, p. 185).

The epigraph above depicts an increasingly dominant vision of medicine for the future. To create a more comprehensive understanding of human health and disease,

L.F. Hogle (✉)
Department Medical History & Bioethics, University of Wisconsin-Madison, 1142 Medical Sciences Bldg, 1300 University Ave, Madison, WI 53706, USA
e-mail: lfhogle@wisc.edu

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_17

it has been argued, calls for collecting ever-greater volumes of information about individuals, in continuous, real-time streams, and from sources both inside and outside of the clinic. The highly influential U.S. National Academy of Sciences commissioned a report to forward this vision, recommending major changes in the way research and the collection of evidence is currently done (NRC 2011).[1] Recognizing that many non-genetic factors are involved in disease risk, the Committee called for the collection of environmental, behavioral and socioeconomic factors, which might lead to the development of more definitive phenotypes of diseases. This way of thinking about research and clinical practice has been dubbed "Precision Medicine," indicating that many data points can more precisely inform understandings about disease at the individual level, and potentially offer treatments more specific to disease subtypes.

The "billions of data points" described above would thus include information about individuals from biorepositories, from their medical histories, and from additional sources thought to provide insight into exposures and interactions with individuals' environments. The tools involved, next-gen sequencing, imaging, real-time streaming and more, produce large volumes of data. Additional sources not previously considered to be "medical" would also be used, including social media interactions (tweets, Facebook or other posts), consumer purchase patterns, internet searches, data streamed by consumer digital devices and more Schadt 2012). Such data is heterogeneous, unstructured, and high-dimensional, requiring the use of analytics that can handle very large, complex data sets, run on massively parallel software run on many servers. Precision medicine is a paragon of big data medicine.

However, in order to work, information would have to be both interconnected and accessible across institutions, regions, and data type by a wide variety of researchers, something that would be difficult to achieve with existing data architectures and ways of collecting and storing information. To this end, the NRC report calls for the creation of a "Knowledge Commons," a new infrastructure for data collection and sharing. A Commons would require "... a massive reorientation of the information systems on which researchers and health-care providers depend" (p. 55). A major part of the reorientation would be to make data acquisition a part of routine clinical care, making each clinical encounter a research-gathering event. Then, the information gathered has to be converted to useable and transportable data, and data flow must be facilitated so it gets to the right users without falling into the wrong hands. To accomplish this, the architects of precision medicine envision "the elimination of institutional, cultural and regulatory barriers to widespread sharing of the molecular profiles and health histories of individuals, while still protecting patients' rights" (ibid, p. 60).

[1]The report followed much earlier similar recommendations from NIH chief Frances Collins developed at the turn of the twenty-first century (Collins 2004). Committee members authoring the report included genome scientists, clinician-researchers in academic medicine, and representatives from Pfizer and a former scientist for Bristol-Meyers Squibb.

Thus, it is not just the tools producing data or the analytics, predictive and encryption algorithms, or cloud storage that are key to big data in biomedicine, but also the legal and social practices of dealing with information about individuals, the management of workflow, organizational policies and procedures, various types of skills and expertise, and the financial system to pay for all of it. In other words, to understand ethical, social and policy issues related to big data in biomedicine, it must be studied as a socio-technological system (Law and Bijker 1992). Technologies succeed—or fail—to the extent that such systems are developed to support them. This may entail enlisting the participation of key organizations and opinion leaders, framing initiatives in a way that aligns them with popular ideas or political currents, creating or removing laws, guidelines and protocols to establish authority, redirecting funding, and more.

In this chapter, I highlight the way that the infrastructures are being built to support big data in biomedical science and clinical medicine as a socio-technical system.[2] I focus in particular on legislative and institutional actions in the U.S. that create conditions of possibility for big data-related technologies to flourish in health and medicine. The form that the big data socio-technical system ultimately takes will likely create new ways of producing and organizing knowledge, new structures to manage the flow of information, and a reordering of relations. Indeed, new organizational forms and social alliances are already emerging in response to new technologies and changes in policy. Such arrangements and investments serve to establish legitimacy and credibility for precision medicine-oriented approaches, as big data projects begin to take on a taken-for-granted presence. In the process, participants must negotiate the continuities and discontinuities with existing policies and procedures, including ethical guidelines. Illuminating these processes draws attention to the less-visible infrastructures being built around big data, which are certain to alter relations among clinicians, patients, commercial entities, payers, researchers, and the State, and ultimately will affect the way we think about human disorders and how to deal with them.

2 Analyzing the Ethics and Politics of Big Data Infrastructures in Biomedicine

2.1 The Scope of Big Data Use in Biomedicine

The emphasis in biomedicine over the past couple of decades has been on genome-related factors in disease. Yet it is increasingly clear that genome variations alone cannot explain the way complex diseases manifest, or in whom. Investigators are using data-driven techniques to identify associations among many variables that

[2]My data comes from field notes and interviews at multiple conferences focusing on big data and biomedicine since 2013, as well as content analyses of policy documents.

may help in understanding disease risk and identify biomarkers that can be used to predict who may become ill and when. Some applications of big data in biomedical research and clinical decision-making have produced promising results (Halamka 2014; Kohane 2011; Sulzicki et al. 2012; Xu et al. 2011; Weiss et al. 2012). While most of these use already-existing data sets, investigators argue that they need access to much larger cohorts of both disease-related and healthy normal individuals, stimulating calls for population-wide recruitment of additional subjects to volunteer their data.[3]

There are other reasons to collect such vast troves of information on individuals besides "personalized" approaches to medicine, having to do with cost containment, medical error mitigation (risk management) and other operational and outcomes research. There is considerable interest in aggregating payer claims data with clinic operations records, prescribing patterns, decision support systems, and a range of social and behavioral data. Big data techniques are used to look for associations indicating adverse reactions, adherence (or non-adherence) to therapeutic regimens, and predict risk of further illness, in particular, readmission to a hospital or other additional cost burdens. Insurance companies, employers providing health plans for employees, and government policy makers are very interested in this information as well.

The ethical and social issues become apparent with such uses. Linked data can be used to stratify individuals into new types of risk or cost pools, which may have implications for patients in a system where care is not universally covered. Concerns about privacy and data security are the most frequently raised concerns about big data and the use of such a broad capture of information about individuals, and certainly, the collection of digital traces also smacks of Orwellian surveillance and exploitation (Kahn 2013; Lupton 2014; Pasquale and Ragone 2014). Considerable effort is being devoted to procedural and technical fixes that provide a sense of protection and trust, while still enabling access to information (Malin et al. 2010; Fredrikson et al. 2014; Kaye 2015). Yet I worry that a preoccupation with procedural matters of privacy can result in a sort of perceptual blindness; that is, attention may be drawn away from some fundamental shifts occurring with potentially far more profound ethical implications. Many discussions about privacy assume that the concept means the same thing to patients, data scientists and bioethicists, and ignore the more subtle ways through which protections (often presumed to be dispensed with through informed consent protocols) are being institutionally changed, resulting in different *practices* around privacy and information. It is important to look beyond governance issues related to privacy alone, as Mittelstadt and Floridi have noted (2016).

The hoped-for power of big data in biomedicine is predicated on the assumption that big data techniques will produce clinically useful and actionable knowledge,

[3]Recent national initiatives to facilitate such large-scale research include, eMERGE (an NIH-sponsored consortium to link data from DNA biorepositories with electronic medical records) and dbGAP (NIH-sponsored database on genotypes and phenotypes) (McGuire et al. 2011).

superior to existing epidemiological and operations research methods, or at least as a powerful and cost-effective adjunct. Yet big data technologies can only acquire credibility and authority with the ability to convince scientists, policy makers and clinicians that big data is a valid way of producing knowledge (but see Neff 2013). After all, it is a radically different way of thinking about research and the associations among variables and predict risk (Cohen et al. 2014; Singer 2014). At the same time, political authorities must be convinced by other utilities; namely, that big data can produce cost savings to a burdened health care system and a viable means through which products may be introduced into the economy. Such broad-scale programmatic initiatives must be authoritative to various publics, including disease advocacy groups, groups interested in taxpayer rights or 'smaller' government footprints, and more. To this end, new infrastructure initiatives use rhetorical framings of "freeing" the data to flow and making medicine more "patient-centered." These discursive parts of the socio-technical system are important in persuading political leaders to support policy instantiating big data approaches (Star and Ruhleder 1996).

Finally, the ethics of big data techniques and practices cannot be understood in isolation. Examining data security protocols or privacy procedures may provide some insight into ethical, legal and social issues and how they may be resolved, and studying the content and development of algorithms is important in highlighting how knowledge gets produced in particular ways, including what (and who) is left out of calculations and predictions that may affect millions of people. Yet alone, these are insufficient to provide a broader understanding of how we came to accept data-driven epistemologies as valid forms of evidence, or how technologies are co-produced in interaction with broader political, economic and social conditions. The conditions of possibility that enable certain technologies to flourish (or not) may include historical precedents, particular political debates or controversies, or as Sheila Jasanoff argues, different modes of public reasoning—what she terms "civic epistemologies" (Jasanoff 2005). This includes what comes to count as evidence in different societies, and how evidence is drawn upon and framed to produce public policy. In the U.S. case, a central point is that in contrast to some other societies, the U.S. does not have a long history of digitized medical records, nor centralized databases or ways of linking medical with other civil databases. Also significant in the U.S. context is the existence of private payer plans rather than single-payer systems, and the long political battles over health care reform.

2.2 Why Study Infrastructure?

Infrastructures can be taken to mean the technical and organizational systems, facilities and tools needed to make things work. When successful, infrastructures become a 'naturalized' part of everyday life and the work and politics of coordination, routinization, standards-making, sticking points and challenges disappear (Edwards 2003; Edwards et al. 2009; Star and Ruhleder 1996). With a successful information

infrastructure, data would move across platforms, organizations, disciplines and institutional environments without friction. Yet much has to happen to get to this point. Scrutinizing the points in development where infrastructures were constrained or made to flow freely can reveal a good deal about political-economic contexts in which technologies arise or stumble, and tensions between prospective stakeholders and users. Infrastructure studies also examine how facts and theories are circulated and mobilized; how they gain traction and the power dynamics involved.

By examining the work of infrastructure-making in action, medicine, legal, ethical and social issues related to big data can be better understood. The technical tools to make the vision of big data in biomedicine thinkable exist, but data has to be amenable to being circulated and has to be "algorithm-ready." In particular, medical records would need to be digital, become more standardized, and more interoperable in order to be transmitted across institutions, to be decipherable across databases, and to serve the variety of repurposing that has been proposed. This involves mundane acts of standardization, but also major institutional changes. In the U.S., this includes converting paper and document-based records to digital, electronically based systems, which has been slower than in many countries. There are many additional technical hurdles, including making the interfaces between data sets and devices interoperable (for example, application programming interfaces (APIs) for software are vendor-specific, made over different times for differing purposes). There are also issues with verifying algorithms, creating encryption and other security measures, and more. More fundamental is the need for a structure within which to exchange information—and incentives for people to use it. For hospitals or researchers to invest time and resources in collecting and disseminating information beyond their primary goal of patient care or a specific research aim (often partly defined by a funder) may take more than a generalized belief that shared information would provide some hoped-for public benefit. The large investments necessary to collect genome and other information as a matter of routine care is a big 'ask' for an economically stressed healthcare system when much of the information may be for long-term research purposes and to build a knowledge 'network,' and may not have direct clinical relevance. It would be disruptive to workflow patterns and traditional interactions among patients, caregivers, providers and researchers. In particular, precision medicine as proposed will blur research and clinical domains, which medicine has tried to keep separate, while further blur boundaries between medicine and commercial consumer worlds (Hogle forthcoming). It would require large investments at the national level that might otherwise go to other public health needs, and at the individual clinic level, could instead go toward specific patient services. Choices made about such matters are thus simultaneously technological, political and moral.

Some of the proposed big data approaches may never become feasible or clinically actionable, yet the infrastructures currently being put into place may endure, affecting research and clinical practice for years to come. The way infrastructures initially develop is consequential: as infrastructures are taken up and stabilized, other ways of thinking and doing are closed off (Edwards 2003). Still, concepts and interpretations also evolve, including competing notions of ethical principles and conventional concepts of protections.

This will become clear in the following sections, where I describe laws that are emerging to govern health information technology (IT) which mandate new organizational forms, make new resource commitments, and create new ways of thinking about evidence. I also include changes to regulatory law as well as fresh changes to the Common Rule governing protection of human subjects. This will be followed by a brief description of other initiatives embedded with big data epistemologies.

3 Setting the Stage for Big Data in Biomedicine: A Brief History

The few existing accounts of big data in the biosciences and medicine suggest that genome technologies, especially high-throughput techniques, are responsible for the meteoric rise in adoption of data-driven approaches. While these technologies accelerated their uptake, it is important to keep in mind the historical roots of health information technologies that shape current conditions. While it is beyond the scope of this chapter to relate the story in detail, it is worth noting the political histories in which contemporary practices are grounded. Medicine was slow to adopt computer technologies, which had developed largely in the military in the 1950s (November 2012; Stevens 2013). Physicians resisted, and policymakers were loathe to spend taxpayer money to support it as a federal program, Even where adopted, software and hardware systems within institutions were often cobbled together by different vendors, with little interoperability within departments, much less across institutions. These origins resulted in problems of being able to get needed information about patients between clinics, or even between departments within a clinic, and investigators had to create new data sets de novo with each new major research program.

By 1961, the Healthcare Information and Management Systems Society (HIMSS) was established, both as a professional association for the fledgling field and as an industry lobbying organization to promote the use of information technologies in healthcare organizations.[4] Over the years, HIMSS advocated for a federal mandate to adopt electronic medical records, ultimately persuading President Bush to develop a national policy on health IT using an economic argument, promising that health IT could save the country $80 billion annually. As a result, a national plan promoting the broad adoption of IT in health was included in the President's "Promoting Innovation and Competitiveness" agenda. A new Office of National Health Information Technology (ONC) was created in 2004 and $100 million was earmarked for demonstration projects. Yet, again there was strong resistance to investing in a national organizational structure, especially in an era of downsizing government initiatives (Brailer 2009). With a new congressional

[4]HIMSS remains a powerful lobbying force and currently has 57,000 individual members internationally, plus more than 615 corporate members and 400 not-for-profit partner organizations.

administration in office, this was reversed when President Obama identified health IT as a national priority, again emphasizing the potential for cost savings and efficiencies. Between these two administrations, health IT became one of the largest publicly funded infrastructure investments the U.S. has ever made.

Developing a health IT infrastructure was no small undertaking. In addition to enforcing the transition to electronic health records, creating standards and ensuring data security, the ONC also had to provide a governance structure for newly-created health information exchanges (HIE).[5] These are collaborations to manage electronic exchanges of information among clinics, care providers, pharmacies, labs etc. with data sharing agreements. They can be regional or state-based or can be a collection of health providers (public or private) or even software vendors establishing an exchange for their clients. National e-Health Initiative was created as a public-private collaboration to consider criteria for certification and interoperability standards for vendors and users (later subsumed under industry-driven HIMSS) and the Federal Health Architecture was created to coordinate health IT actions across 20 federal agencies.

Tracing the roots of health IT makes the politics more visible, Large-scale efforts to create infrastructures to accommodate digitization of medicine and big data analytics are as much about operations efficiency and cost savings as they are about quality of care and personalized approaches, as is often claimed. I turn next to legal infrastructures.

4 Legislating Digitization and Data Sharing

The most significant legislative actions in the U.S. setting the stage for big data in biomedicine are the Health Information Technology for Economic and Clinical Health Act (HITECH) passed in 2009 and the Affordable Care Act in 2010.[6] Provisions in these acts changed the way many services are financed in the U.S., established new institutions and organizational structures, and promulgate additional related technologies.

4.1 The Health Information Technology for Economic and Clinical Health Act

HITECH mandated the continuance of the ONC, and established incentives and disincentives to participating in a national health information infrastructure. Pres-

[5]Electronic health records should be distinguished from medical records. The latter is essentially the entirety of what would appear in a patient's paper chart from a single or primary care provider, including physician notations, tests and treatment history. The former is being used by the ONC and refers to all information from all care providers.

[6]Pub.L. 111–5.

ident Obama used the venue of the American Recovery and Reinvestment Act of 2009 (ARRA), his economic stimulus initiative, as the venue for passing HITECH, indicating that economic rationale was again as much the selling point as health improvement.

The Act authorizes the Centers for Medicare and Medicaid Services (CMS, the federal venue for paying for the elderly and young children; it also oversees the Health Information Portability and Accountability Act) to pay incentives to clinics and physicians adopt EHRs. However, in order to receive incentives, care providers must meet specified actions and practices that indicate that health information technologies are being meaningfully used for the benefit of patients and health improvement. So far, about 800 hospitals and 33,000 physicians have been paid under the plan. The Act provided for $32.7 billion in incentive payments and about $2 billion in grants for programs to instantiate its provisions (Redhead 2009). About $28 billion has been spent to date. State governments also received about $560 million in funding to stimulate the uptake of HIEs (Blumenthal 2010).

The HITECH Act first incentivizes, then penalizes healthcare facilities that do not adopt EHR through these "meaningful use" provisions. That is, there are three stages of increasingly complex requirements that must be met and for which providers can receive payments, then after 2015, the incentives disappear and CMS can also reduce its usual payments for services to providers which have not met the requirements. Precision medicine goals are built in to the requirements: for example, requirements call for ways of providing patients with a way to directly access their health information, through portals, apps, or direct messaging.

Stage 3 provides a greater role for user-generated data, largely from mobile health devices, including so-called "BYOD" (bring your own device—smart phone or tablet), wearable sensors and more, to produce data. This meshes with new regulations allowing the use of such devices for regulatory oversight and research (see Sect. 6).

4.2 The Affordable Care Act: Building Infrastructures in to Health Care Reform

The Affordable Care Act (ACA) of 2010 was the major legislation reforming the way health care if provided and paid for in the U.S. It also has several provisions for data infrastructures.[7] The establishment of "Accountable Care Organizations (ACOs)," for example, creates a new payment model that rewards health care providers (those receiving Medicare & Medicaid Services (CMS) reimburse-

[7]The formal name is the Patient Protection and Affordable Care Act, PL 111–148 (2010), more commonly referred to as the Affordable Care Act (or colloquially, "Obamacare"). It is important to distinguish that unlike many European countries, Americans are still provided health care insurance through their employers, through private exchanges, rather than being provided to all by the State, with the exception of the elderly, poor, and some children. This Act was an effort to provide more Americans with coverage.

ment) based on demonstrated health outcomes of patients using selected measures. These measures require integrated data systems to track individuals and cohorts of patients. The system also ties these data to cost data via insurance claims, prescription orders, clinic operations costs and more. Such linkage is a major benefit for payers, enabling access to information that may help them in the stratification of patients according to risk and cost. Providers are rewarded with financial incentives when they meet specific criteria for quality, which includes cost and medical error reduction, and lowered costs for defined populations. ACO's are then held accountable for achieving quality improvements according to these measures and reducing costs. For example, CMS is authorized to reduce payments to hospitals with excess readmissions (e.g., when a patient returns to the hospital within 30 days of discharge).[8]

One such quality improvement measure—adherence—is significant because it has spawned considerable activity from big data entrepreneurs, payers, and pharmaceutical companies. Non-adherence to prescribed drugs and treatments can result in unabated chronic symptoms and thus can be costly to health plans (Sulzicki et al. 2012). Big data tracking of prescription purchases (or failure to fill or refill prescriptions) is being done by linking EHR with pharmacy data, as well as social media tracking. For example, Optum (an arm of the for-profit insurance provider United-Health) uses predictive analytics to identify individuals who "at risk" for negative adherence behaviors.[9] Patients who are noncompliant then get regular reminders to take their prescriptions, but also are profiled as "noncomply-ers" or "health deny-ers" on self-efficacy scores by payers and providers, altering their identity as high risk, high cost patients. Drug adherence risk scores can be calculated for all members of a health plan, which can help providers meet required quality measures.

Another ACA provision requires hospitals and physicians to provide patients with a way to view and download their medical data (Ricciardi et al. 2013). While intended to provide transparency for patients to see their data for their own use, another click provides the easy capability to transmit all their medical data to third parties (including research studies and other entities).[10] The technology is framed as being "patient-centered," because it "empowers" patients to become involved in their own care. This capability is consistent with Precision Medicine goals to

[8]Section 3025 of the Affordable Care Act added section 1886(q) to the Social Security Act establishing the Hospital Readmissions Reduction Program.

[9]Craig Schilling (developer of Drug Adherence Index™ Optum) presentation to Medical Information World, Boston, May 2014. As Schilling explained, it is too costly to target all patients for compliance; the algorithm helps to identify which potential problem patients to target. The ACA quality measures are on a star rating, and as Schilling put it in his presentation: "at the end of the day, it's all about the star rating."

[10]The Office of the National Coordinator and the White House are encouraging IT developers to develop standards for automatically uploading new data, and for stimulating new markets around this activity (ONC 2015a; Ricciardi et al. 2013).

increase the number of people freely uploading personal health information, but as I describe below, this information transfer relies on data security measures and is not in a legally protected space for privacy.

4.3 Infrastructure and Meaningful Use, or Excessive Regulation and Boondoggle?

The HITECH meaningful use requirements and ACA quality measures requirements have been fraught with political and technical problems, begging the question of whether the requirements and incentives made enough difference to "free the data" available for big data analysis, and if the enormous investments made a difference in health outcomes.

To encourage the development of HIEs, a considerable amount of federal funding has been provided in grants and other financial incentives. The CMS provides 90 % matching funds for HIE activities and the ONC supplied a round of $60 million in grants for 64 projects on issues such as network design, data security, and policy for secondary use of data through the Strategic Health IT Advanced Research Projects (SHARP) Program. A State Innovation Models Initiative has provided about $1 billion to help states develop and test models for exchanges, and CMS Health Care Innovation Awards also award up to $1 billion to test new models of service delivery using electronic infrastructures. The large investments have raised questions for some observers who argue that a few entities are profiting without much overall positive outcome (Thune et al. 2015; Adler-Milstein 2013). At the time of the first ONC progress report, the number of hospitals using electronic health records increased significantly, but sharing information was a different matter. Despite the financial incentives, fewer than 30 % of hospitals and only about 10 % of ambulatory care facilities participated in an HIE by 2013 (Charles et al. 2014). Of those hospitals exchanging information, more than half were exchanging only some information (such as lab tests), and were often not sharing outside their own healthcare system. Fewer than half notified physicians if one of their patients was in the hospital, and only about half were able to send and receive patient health information from outside their system electronically. Only about 10 % had a system that allows patients to share their health information with a third party—an area of intensifying efforts of big data users (ONC 2014).

The funds for incentives are disappearing, and the current models for HIEs that rely heavily on grants are not financially sustainable. Significantly, private HIEs (often operated by commercial vendors) charge varying (and sometimes high) fees to participate or to access data sets, raising ethical questions about limiting access (Vaidya 2014). Pragmatically, there are simply no incentives for hospitals to pay for expensive infrastructures out of their own budgets to pay for interfaces to exchange information, or for designing systems that will not capture revenue, such as systems to share data with patients (Adler-Milstein 2013). The state and regional HIEs do

little to connect information at a national level, plus small clinics and many post-acute facilities, including rehabilitation and long-term care facilities, are not eligible for incentives and do not participate in HIEs.

The vision for health IT failed to take into account such realities, as well as the lack of interoperability of IT platforms (Adler-Milstein and Jha 2013). A commissioned report highlighted the interoperability issues, noting that different federal agencies have authority over different aspects of EHR implementation and exchange, but there does not appear to be any single interagency group charged with coordinating and harmonizing these efforts (Agency for Healthcare Research and Quality 2014, p. 15).[11] This state of affairs makes the importance of infrastructure more visible; it is when things do *not* function as intended that the frictions between old and new practices become evident.

5 Considerations of Privacy and Protections

Recognition of a concern about privacy and data security are a consistent theme in documents directing the building of new information infrastructures. Yet provisions in new laws create procedural and definitional moves to get around the problem without addressing directly what privacy is in an era when scientists are urged to have open-source data sharing, when algorithms make it relatively easy to re-identify anonymized data, and when individuals volunteer considerable information about themselves in social media and other venues. Policy makers are making directives about how to manage information that enables information to flow in desired directions, yet may conflict with existing guidelines and understandings of what constitute appropriate protections.

5.1 Provisions in HITECH and Existing HIPAA Rules

The HITECH Act recognized potential problems with data privacy and security issues, making provisions for protecting against commercial use of EHR, and adds criminal penalties for breaches to privacy under the Health Insurance Portability and Accountability Act (HIPAA) (Blumenthal 2010). HIPAA, enacted in 1996, authorizes the Department of Health and Human Services (HHS) to create regulatory rules related to the privacy of individually identifiable health information, particular

[11]The JASON Report cited ineffectiveness of the Federal Advisory Committees created to assist with coordination (the Health IT Policy Committee and the Health IT Standards Committee, which report to the ONC) (Agency for Healthcare Research and Quality 2014). Notably, a proposed new law would replace existing committees with a new entity comprised of industry representatives (see Sect. 7).

with regard to unauthorized or inadvertent disclosures.[12] Much has been written about HIPAA, including its weaknesses, but for the purposes of this chapter, a few main points bear highlighting: (1) while it covers information transmitted between health care providers and relevant "covered entities" and business associates, these are ambiguously defined, and (2) it covers the transmission of information between care providers and business associates, but not individuals, and not the manner of collection.[13] Importantly, any information shared by individuals—whether it entails patient experience narratives uploaded in a survey, medical record data voluntarily shared by individuals, posted in social media or elsewhere, is not HIPAA protected. Furthermore, as other chapters in this volume will attest, the ability to re-identify individuals using diagnostic codes or other relatively easily accessible publicly available information and re-identification algorithms means that there can be no guarantee of anonymity (Loukides et al. 2010; Gymrek et al. 2013).

Nicholas Terry argues that these situations create an especially problematic situation with the introduction of specific policy initiatives to promote broader collection and sharing of information and the employment of big data (Terry 2012). He points out that our "medical selves" exist outside of HIPAA-protected spaces, both in terms of what individuals make available themselves, and what data aggregators and brokers are able to infer through data mining of social media entries, purchase patterns, prescription use, sensor data and much more. The surrogate selves represented by aggregating and triangulating such data mean that no protected health information may need to be accessed in order to infer behavior, illness or other characteristics of individuals. With such end runs around privacy protections, the policy response has been to rely on informed consent to attempt to preserve autonomy. As Barocas and Nissenbaum (2014) demonstrate, this is not only insufficient, but the use of big data, particularly with predictive analytics and the ability to infer properties across groups, creates complex problems. Even if only a few people volunteer information, it implicates many others.

[12]Health Insurance Portability and Accountability Act of 1996 Public Law 104–191, Sub. F, Sec. 264.

[13]Health information is defined as "any information, whether oral or recorded in any form or medium, that (A) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (B) relates to the past, present, or future physical or mental health or condition of an individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual." Personally identifiable information is that which can be directly tied to an individual, including name, geographic information smaller than a state, social security number, birth and death dates, phone and fax numbers, device serial numbers, and biometric identifiers (including voice print and photos) (45.CFR 160.103). "Business Associate" is defined in 45 CFR 160, subpart A and includes an entity which " . . . claims processing or administration, data analysis, processing or administration, utilization review, quality assurance, patient safety activities, billing, benefit management, practice management and repricing, or provides legal, actuarial, accounting, consulting, data aggregation, management, administrative, accreditation, or financial services to a covered entity or for an organized health care arrangement . . . " A summary of the privacy rule can be found at http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html

However, it is also the case that investigators increasingly want to use explicitly identifiable information in order to link 'omics' data with visual or audio data. In vitro disease modeling, for example, may seek to link what they see in the dish with in vivo behavioral phenotypes (Saha and Hogle 2014). This sets up a growing ethical tension between the need to specifically identify individuals and to protect their identity (Kaye 2015).

There is also little to prevent non-covered entities (or covered entities acting for a different purpose than outlined in HIPAA) from sharing data. For example, private HIEs, investigators and other holders of data sets may release data under restrictive or generous data-use terms, or may repackage and sell the data to third parties. If data holders such as an NIH-funded project, a private or public biobank, etc. dissolve, they can transfer or sell entire data sets, which would most likely not have been covered in original consent agreements. In sum, HITECH prioritizes privacy and data security, but neither HITECH nor HIPAA solve some of the actual problems, including tensions between the need to access data and the need to protect human subjects (see also Kaye 2015).

There are some prohibitions on the sale of protected information through HIPAA.[14] However, the Center for Medicare and Medicaid Services (CMS) which also has data on claims for services, personal health information, and identified providers, recently reversed its policy prohibiting access to CMS data for commercial use (e.g., to develop products or tools to sell on the market) to allow such uses in the interest of providing data for the 'public good.' Individual records will be de-identified, but the identity of care providers will be available. As Niall Brennan, CMS chief data officer and director of the Office of Enterprise and Data Analytics put it, "as the [health care] delivery system transforms *from rewarding volume to value*, data will play a key role" (emphasis added) (CMS Press release, June 2, 2015).[15] This underscores the argument of economists, management consultants and entrepreneurs that health data has value as a product in its own right.

Finally, the President's Council of Advisors on Science and Technology (PCAST), an authoritative scientific advisory body, published a report dealing with big data privacy issues, identifying interception, stalking, false or spurious facts constituting misleading profiles of individuals, and loss of autonomy as serious threats to privacy (President's Council 2014). However, it focused on downstream effects (after a harm might have been incurred), and advocates the use of tools such as consumer "preference profiles" in place of or as an adjunct to formal consent processes. Recognizing the increasingly invasive practices of data brokers and aggregators, the White House also issued a report, advancing a previously-prepared Consumer Privacy Bill of Rights, but did little more than calling for data brokers to be more transparent about their practices of gathering and using data on individuals (Executive Office 2014).

[14] Id at 164.508 'uses and disclosures for which an authorization is required.'

[15] https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2015-Press-releases-items/2015-06-02.html

5.2 *Changing the Common Rule: What Is a Human Subject?*

At the same time as modifications to privacy protections are taking place, the U.S. is in the process of making major changes to the Common Rule, the major federal regulation for the protection of human subjects that applies to all federally funded research.[16] An Advance Notice of Proposed Rule-Making (ANPRM) was published in 2011 with the title giving a clue about the reasons for change: "Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators."[17] The proposed rule explicitly names new technologies as a chief reason for the needed update, including rapid genome sequencing, imaging, and data analytics, indicating an awareness of the growing use of both big data techniques and re-identification algorithms.

After almost 4 years of deliberations and many thousands of comments from supporters and detractors, the revised Notice was published on September 8, 2015.[18] The 519-page document proposes to strengthen data security measures, centralize IRB oversight for multi-institution studies, make stricter requirements for what must be included in consent forms, and makes recommendations for standardizing informed consent procedures. Concern about genetic privacy prompted a change requiring patients to consent for each use of biospecimens, even if it had been de-identified and had previously been used for studies and was stored, or was left over after being used for clinical purposes (e.g., a blood sample taken for diagnostic testing).[19] However, in contradiction, the new rule states that broad consent should be used; that is, participants should agree to allow secondary uses of their material and information beyond the original study without having to be re-consented. This was in response to a concern expressed by investigators about the perceived increased burden by having to consent for each use.

[16]The Common Rule (45 CFR 46) was created to implement uniform regulations across the major federal agencies, including the Department of Veterans Affairs, Environmental Protection Agency, National Science Foundation, Agency for International Development, Department of Defense, Department of Commerce, Department of Education, among others. Researchers funded by these agencies are subject to the Rule. The proposed rule extends the scope to non-federally funded studies as well.

[17]The ANPRM and public comments can be found at http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html and in the Federal Register at 76 FR 44512–44531. The NPRM will be available online at http://federalregister.gov/a/2015-21756

[18]In the process of conducting research on this topic, I attended a number of medical information technology and precision medicine symposia and workshops, and in each one the desire to change the Common Rule was raised in presentations and audience comments. In informal discussions with participants as well as public presentations sponsored by the White House, it was consistently asserted that this was a top priority, and that it would happen by September.

[19]Specimens that have been stripped of identifiers ("de-identified") are currently not counted as human subjects. Currently, the definition of a "human subject" includes living subjects about whom a researcher obtains data through intervention or interaction with the individual, but also any biospecimen or data derived from a human and for which any individual personal information can be identified.

More notable for big data researchers is the creation of new categories of research exempted from IRB review. This includes studies with "informational risk that is no more than minimal" which is now defined as: use of data containing personal health information originally collected as a part of a non-research activity (assuming a notice of the possibility of such use was given), and public behaviors or activities common in everyday life when sensitive information may be collected (provided that data security protection procedures are followed), or focus groups and surveys.[20] In so doing, the new rules remove oversight for a great many of the kinds of studies emerging with big data that might involve informational risks, stating: "IRB review or oversight of research posing informational risks may not be the best way to minimize the informational risks associated with data on human subjects. Instead, informational risks may be best mitigated through compliance with stringent standards for data security and information protection that are effectively enforced through mechanisms such as periodic random audits." Jurisdiction for such risks is pushed back under HIPAA, which as stated previously, only covers certain kinds of conditions, and while protecting the unauthorized access to personal health information, does nothing about the *collection* of data. Also, HIPAA has historically not allowed broad consent, setting up yet another potential conflict between mandates.

This represents a major change in the current scheme in order to concentrate oversight on higher-risk research, but also opens the door for many of the sources of data that will be used in big data research on individuals. Furthermore, with the broad use of algorithms to create shadow or surrogate identities from digital exhaust the question of what counts as a human 'subject' may be up for grabs (Terry 2012). The deliberation and ultimate establishment of new human subjects rules clearly demonstrate the co-production of science and society in the context of increasing tensions around how societies have thought about protections of human subjects and imperatives to have unfettered access to medical and scientific data.

6 Regulating Products Using Big Data: Changes in the Food and Drug Administration

Data-driven science is affecting regulatory oversight as well as discovery research and clinical medicine. There are noteworthy ethical implications for the production of evidence used to support product reviews, the scope of what may be regulated, and a statutory change which allows the FDA to have direct access to patient information without their knowledge or consent. While most of these changes are in response to calls to streamline regulatory processes and reduce barriers to new

[20]This is based on the assumption that people engaging in activities occurring in a public context would have no reasonable expectation of privacy.

product entry, they raise new questions about what is in the public's health interest and how it is best served by public agencies.

The FDA has begun to encourage the use of electronic health records in clinical trials, both to identify and recruit potential trial subjects more efficiently and less expensively, and to mine the data for non-interventional clinical trials. Such observational studies retrospective mining of existing, de-identified records, sometimes called "covert" trials, because no additional consent is required: no intervention is being made and de-identified records would not be counted as human subjects. To illustrate, some researchers are following outcomes of groups of cardiovascular patients who have had different interventions for the same condition. Cohorts of patients in a hospital's records database using one treatment could be compared against others with different treatments, or against healthy controls also in the database, or against patients from other hospitals' databases. This method requires very large numbers of patients, and may entail data use agreements with other institutions (field notes June 6, 2013).

There are intense disagreements about how to regulate devices and software that will be used to produce the volumes of information used in big data analyses,[21] The explosion of various mobile health devices and applications—including those intended to monitor disorders (e.g., wearable sensors paired with automated detection algorithms for arrhythmias), to track triggers and use (or non-use) of medications (e.g., GPS tracking devices on asthma inhalers) as well as the many 'consumer' health devices (e.g., fitness, sleep or nutrition monitors) have further blurred definitions of "medical device" and "consumer health product," as well as definitions of users as "patient" and "consumer." This has created quandaries for regulatory and legal purposes, since some uses of such devices and apps for smart phones and tablets may be used for diagnosis and potential treatment decisions, while others are often cast in a framework of consumer sports and entertainment.

A report from the Institute of Medicine (IOM) (commissioned by the ONC) advised that the FDA was not the appropriate agency to oversee "health" IT, and that such regulation would hinder innovation. The authors strongly endorsed the use of information generated from mobile devices in clinical decision-making and argued that "only the minimal set of standards or requirements necessary for key functional utility [should be specified]." The report framed digital health products as patient safety tools, and advocated the "emergence of a digital infrastructure that allows data collected during activities in various settings – clinical, research, and public health – to be integrated, analyzed, and broadly applied to inform and improve clinical care decisions, promote patient education and self-management,

[21] Guidance document: Electronic Source Data in Clinical Investigations, available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm328691.pdf. The FDA defines an electronic record as any combination of text, graphics, data, audio, pictorial, or other information represented in digital form that is created, modified, maintained, archived, retrieved, or distributed by a computer system (21 CFR 11.3(b)(6)). Source data includes clinical findings, observations, or other activities in a clinical investigation used for reconstructing and evaluating an investigation.

design public health strategies, and support research and knowledge development efforts in a timely manner." That is, information should be collected once, and then used for multiple purposes.[22]

At stake is what kinds of information might require interpretation from a medical professional, or where the app might be used in determining treatment or mitigation of conditions, particularly life-threatening ones (glucose monitoring for diabetics, or even software interpreting the results of laboratory tests), and when the user might be dependent on the apps for this information. Professionals in clinical decision support fields and laboratory sciences oppose deregulation, while many IT professionals and entrepreneurs strongly favored it. Definitive direction for regulation was not established, despite Congressional debate in 2011. The debate was not just about regulatory authority, however; redefining some products as "health products" not "medical devices" places them in a different category for tax purposes (The Affordable Care Act imposed a new tax on medical devices).

A bill entitled the Sensible Oversight For Technology Which Advances Regulatory Efficiency (SOFTWARE) Act (HR3303) was then introduced into legislation. Amended in 2014, the bill would have classified software into distinct categories (health and medical), restricting the FDA's authority to medical alone.[23] Most of the apps used for self-reporting or for aggregating data into large data sets would not be regulated under the proposed definitions. Providing input into the process, the Food and Drug Administration Safety Innovation Act (FDASIA) Workgroup, comprised of academic and industry experts, proposes recommendations for a risk-based regulatory framework while promoting innovation and avoiding regulatory duplication.[24] While the bill was not enacted, much of the language was incorporated into the subsequent 21st Century Cures Act, which makes sweeping changes to regulatory processes (see Sect. 7). In the meantime, the FDA has issued its own guidance (updated in 2015), retaining the discretionary right to regulate software that might be classified as "health" under the proposed statute.[25]

[22] Discussion of additional ethical and legal issues related to mobile health devices can be found in Mittelstadt et al. (2014).

[23] "Health" software is defined as that which is intended for use by patients for self-management or self-monitoring of a disease or condition, including management of medications; is intended for use to analyze patient-specific information or other medical information for the purpose of providing general information related to the prevention, diagnosis, prognosis, treatment, cure, monitoring, or management of a disease or condition; is intended for administrative or operational support for financial purposes; is intended for use for use aggregation, conversion, storage, management, retrieval, or transmission of data from a device or other thing. See http://assets.fiercemarkets.net/public/healthit/softwareact1-15draft.pdf for a full description of the amended draft bill.

[24] The Food and Drug Administration Safety Innovation Act (FDASIA) Workgroup provides expert input on relevant issues as identified by the FDA, the ONC and the Federal Communications Commission (FCC), focusing on safety for mobile medical applications. Members include representatives from Intel, Qualcomm, Roche Diagnostics, Practice Fusion and other (mostly large) corporations involved in developing mobile medical apps.

[25] FDA defines relevant regulated entities as whose which match the definition of "device" and that are intended to be used as an accessory to a regulated medical device,

The most remarkable change, however, has to do with interpretations of the FDA's role and responsibilities to act in the public interest, as statutorily mandated by the U.S. Congress. This is seen in the Sentinel Initiative, a change in policy and practice little discussed outside of regulatory communities. The 2014 Initiative revises the way the FDA conducts post-market surveillance for adverse effects of approved drugs. Previously, potential harms were only detected when larger numbers of people beyond clinical trials began to use a drug, and effects were voluntarily reported. Sentinel authorizes direct access to patient records (from clinical systems and disease registries and repositories) and medical billing records from payers, obtained from major insurers, such as Aetna, Anthem, Kaiser, and Humana.[26] To date, prescription medication data on approximately 178 million people has been accessed. A pilot program conducted at a major health services organization claims to have collected 358 million person-years of data that include 4.0 billion prescriptions, 4.1 billion doctor or lab visits and hospital stays, and 42.0 million acute inpatient stays, and significantly, will continue to collect information on all its patients as a routine part of care in its 18 partner organizations (Findlay 2015). This extends surveillance beyond retrospective record review and enables unlimited longitudinal data collection.

Sentinel was intended to be linked to another initiative, the Nationwide Health Information Network (NHIN), to "connect clinicians across the health care system and enable the sharing of data as necessary with public health agencies." The idea was to create a nationwide interoperable system with information useable for a variety of health care and operations purposes. A new infrastructure was thus created giving the FDA authority to collect and in turn release private health data (including data in identifiable form) to other entities, including both private operators and partners of the Sentinel System and outside data users, which could be academic research institutions and commercial entities.

While new regulations to protect privacy could have been established in the new infrastructure, Congress instead ordered the FDA to comply with existing HIPAA regulations, which allow personal information to be distributed to other parties in circumstances where it is considered to be in the best interest of public health. In such cases, a patient's authorization is not required before information is released to others. The FDA's mandate to protect the public from dangerous products appears to fall within this exception.[27] The FDA has historically never waived the right to consent for product sponsors (except in the trial of an experimental treatment in a

---

or transform a mobile platform into a regulated medical device. Guidance documents can be found at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf. Examples of mobile medical applications that may be regulated can be found at: http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/ConnectedHealth/MobileMedicalApplications/ucm368744.htm

[26] An obvious problem is that medical codes used for mining were designed for billing, and would not necessarily reflect events relevant to adverse event surveillance.

[27] 16 FDAAA § 905(a), 21 U.S.C.A. § 355(k)(3)(C)(i)(III)(aa)–(cc) (West Supp. 2008). Section 164.512 of the Privacy Rule of HIPAA of the Privacy Rule allows "public health" exemptions

life-threatening situation where the person could not reasonably give consent), but with the new structure, it can then determine what meets the definition of a "public interest standard" and disclose information itself without consent.

While the linkage of large databases on citizens is not new in some countries such as Denmark, Sweden and the UK, in the US this is a sea-change. Legal theorist Barbara Evans likens this to other laws governing infrastructure regulation, such as energy or road construction, where there is a governmental claim, say, to a need to have right of way to construct power lines, roads or pipelines. These claims led to laws allowing eminent domain and takings of private property to achieve the public interest in infrastructure.

Evans argues that individuals' statutory rights to protect their privacy constitute a form of transaction cost. As she puts it: "Congress cast medical privacy, a hot-button issue for many members of the American public, as an infrastructure regulatory problem" (2009).

In this way, law governing product review can exercise coercive power, in contrast to their historical practice of relying on voluntary cooperation (e.g., individuals are subject to the new structure, even though ostensibly, they may not be involved in product safety testing). Evans cautions that protocols entailing types of information that may be collected longitudinally should be instituted, with informed guidance of what may be released to external entities.

The Sentinel Initiative has thus become one of several federally-mandated initiatives to establish a national data resource. In the process, it creates another end-run around privacy. The practices and infrastructures developed through this and similar initiatives are likely to be codified in law in the near future. In particular, I turn now to a pending law, the 21st Century Cures Act.

7 Legislating Precision Medicine? The Proposed 21st Century Cures Act

Recently proposed legislation dubbed the "21st Century Cures Act" states as its aim the modernization of discovery, development and delivery of new treatments and therapeutic products. The Act would make substantive changes to the statutory and regulatory framework of the U.S. Department of Health and Human Services (especially the Center for Medicare and Medicaid Services) as well as the Food and Drug Administration (FDA), which regulates drugs, devices and biologics.[28]

in which personally identifiable information may be disseminated without consent from the individuals.

[28] At the time of this writing, the bill (H.R.6) was overwhelmingly approved by the House of Representatives in a rare bipartisan effort (July 2, 2015), and is likely to pass easily in the Senate. Estimated net costs (costs to implement minus projected savings) are pegged at $8.7 billion over 5 years.

Specific provisions are made primarily to move products through regulatory review more quickly, and to enhance research (mainly by increasing funding for the NIH and providing funds for Precision Medicine.)

Big data approaches are embedded throughout the proposed legislation, via explicit calls for data-driven techniques as well as organizational changes to support them. The Act extends HITECH and ACA efforts to create accessible, interoperable databases, but also prioritizes public-private partnerships (cf. Ranck 2014).

7.1 Redefining Research in Order to Access Data

Title One provides funding for data-sharing programs and creates a public-private partnership to facilitate data collection and analysis. Imbedded in this section is an important change to facilitate access to data on individuals. Definitions of research are altered with an amendment to the HITECH Act, such that the use and disclosure of protected health information by a covered entity for research purposes will be treated similarly to those for operations purposes or for the good of public health. This means that while covered entities *may* obtain consent from patients before using or disclosing their information, it would no longer be *necessary*. Furthermore, the original requirements regarding disclosure of personal health information for research purposes does not apply: these include the requirement to get IRB approval prior to use and to get guarantees from the entity or researcher that the use and disclosure is necessary.[29] The Act also calls for the revision of the HIPAA Privacy Rule, giving covered entities and business associates more control over access and dissemination of data without prior authorization, assuming they are taking proper data security measures.[30] Furthermore, they are allowed to collect payments for preparing and sending the data.[31] The Act also directs the HHS Secretary to conduct pilot demonstrations to extend the use of the FDA Sentinel System for adverse event monitoring, and authorizes the FDA to obtain data from more holders through a coordinating center. These demonstrations would be classified as "public health activities," and thus not "research" under HIPAA definitions, thereby further codifying the interpretation that these uses of data are not subject to IRB review. Another provision enables a major departure from existing human subjects protections by allowing studies for which "the proposed clinical testing poses no more than minimal risk" not to have to go through conventional human subjects approvals. Who gets to define "minimal risk" is not precisely specified. In short,

[29]Title One, Subtitle G, Part 4. subsection 13442.

[30]Title One, Subtitle G, sec. 1124.

[31]Subtitle G, Part 4 subsection 13442 states: "(a) Remuneration. The Secretary shall revise or clarify the Rule so that disclosures of protected health information for research purposes are not subject to the limitation on remuneration described in section 164.502(a)(5)(ii)(B)(2)(ii) of part 164."

the Act greatly expands the ability for business entities to use and disclose HIPAA-protected information and conduct research as it is defined in the statute with less oversight.

7.2 New Forms of Evidence

The Act supports the Precision Medicine goals of "eliminating barriers" to research by altering forms of evidence required for regulatory review. Subtitle C, "FDA Advancement of Precision Medicine" modifies the Food, Drug, and Cosmetics Act to add provisions explicitly toward these goals.[32] Here and elsewhere, the Act encourages the use of biomarkers rather than hard clinical endpoints as a measure of effectiveness. Biomarker discovery is an active area of research requiring genome information from very large cohorts of individuals and relying on predictive analytics to find associations (President's Council of Advisors on Science and Technology 2014). Biomarkers may be used, for example, with predictive analytics to identify subsets of patients who may be likely responders to therapies in order to streamline the conduct of clinical trials. It creates a new regulatory class of "precision drug or biological products" which would be evaluated using more targeted trials rather than the gold-standard large randomized trials.

Another Section, entitled "Modern Trial Design and Evidence Development" instructs the FDA to include the use of nonconventional forms of evidence when evaluating new products, including adaptive clinical trial designs, Bayesian probability methods and predictive analytics. Significantly—and controversially—it also encourages the use of observational data instead of relying on the gold standard of randomized clinical trials. Observational data could come from several sources, including data mining of registries, from clinical encounters (patient narratives about their own experience), patient-derived data (from sensing devices or mobile health applications (FitBit, smart phone apps, etc.), surveys, clinical narratives) or clinical outcome assessments.[33] Risk-benefit analyses are altered to allow the use of patient experience data, which the Act defines as: "data collected by patients, parents, caregivers, patient advocacy organizations, disease research foundations, medical researchers, research sponsors or other parties determined appropriate that is intended to facilitate or enhance the Secretary's risk-benefit assessments, including information about the impact of a disease or a therapy on patients' lives."[34] This is consistent with calls from Precision Medicine advocates to recruit

[32] Sec 2014. The relevant FDCA section is at 21 U.S.C. §§ 351–60.

[33] Title Two Subtitle D. Clinical outcome assessments are defined in the Act as "a measurement of a patient's symptoms, overall mental state, or the effects of a disease or condition on how the patient functions" (section 2021). This includes patient-reported outcomes (e.g., "a measurement based on a report from a patient regarding the status of the patient's health condition without amendment or interpretation of the patient's report by a clinician or any other person").

[34] Title Two, Subtitle A, Section 2001.

patients directly and through advocacy organizations to create very large cohorts of subjects willing to volunteer personal health information, as well as allowing access to their medical records.

As mobile health devices are increasingly being used to capture data, the law proposes to accept such patient-generated information as valid evidence to aid drug development and regulatory review. This raises ethical questions for many about whether observational data, along with the associational studies produced by big data, are sufficient to provide evidence of safety and efficacy, since there is less emphasis on causation than correlation. More significantly for the purposes of this chapter are the ethical questions of what comes to count as valid evidence and why? What are the political, economic, and social conditions that enable this radical shift in thinking, and what might be the consequences?

The bill is strongly supported by pharmaceutical, device, and information technology industries.[35] Critics worry about lowering the standards of evidence of safety and efficacy by substituting data produced in such ways for actual outcomes, which have conventionally been measured in carefully controlled clinical trials (Avorn and Kesselheim 2015). They also note the strong lobbying efforts of industry and direct participation in shaping the bill's language and inclusions. For example, the Act grants additional periods of exclusivity for certain brand-name drugs approved for rare diseases or conditions, delays entry for some generic drugs. Title Three also eliminates the existing Health Information Technology Standards Committee and orders that the HHS Secretary contract with commercial accredited standards development organizations. Other critics question the cost structures and effects on overall health care expenditures.[36] Nevertheless, the bill sailed through passage in the House of Representatives, and at the time of this writing, it is projected to pass the Senate (although amendments are likely). Overall, the effects will be to situate industry more centrally in decisions about data infrastructures will allow information to flow and to whom.

[35] Supporters include powerful industry lobbying groups such as PhRMA (pharmaceutical industry association) and Advomed (medical devices industry association) and CHIME (College of Health Info Management Executives).

[36] For example, the bill limits Medicaid payments (health insurance for those with low income) for durable medical equipment (e.g., oxygen tanks, etc.) to states rather than being born by the federal government. So while the Congressional Budget Office estimates that direct spending would be reduced by $11.0 billion (net) from 2016 to 2025, however, the cost to states for Medicaid would be increased $2.6 billion over the same period. Interestingly, funds to pay for the changes would come from the sale of 8 million barrels of oil from the Strategic Petroleum Reserve in each of the fiscal years 2018 to 2025; not surprising since the bill originated in the Energy and Commerce Committee (see http://energycommerce.house.gov/fact-sheet/hr-6-21st-century-cures-act-frequently-asked-questions).

8 Preparing the Way for a Knowledge Commons: Other Organizational and Social Arrangements

Several additional large-scale, publicly and privately funded initiatives have been engaged in building the infrastructures necessary to support the widespread adoption of data-driven medicine. At the White House level, the ONC published the Federal Health IT Strategic Plan 2015–2020, which has several components that explicitly or implicitly favor big data analyses. It promotes the expansion of uses of mobile health devices and other patient-reporting information tools, calls for the broad incorporation of genomic information into longitudinal medical records, promotes the incorporation of medical records and personal health information into clinical trial designs, promotes "analytic capabilities that allow for precision medicine," funds collaborative research data networks and infrastructure, and encourages the collaboration of private with public entities in data sharing. A companion "roadmap" report shifts the aims from adoption of electronic records to addressing the problems of interoperability (ONC 2015a, b see also Grossman et al. 2011).

8.1 NIH Investments in New Research Infrastructures

The National Institutes of Health (NIH) created the initiative "Core Techniques and Technologies for Advancing Big Data Science and Engineering Initiative" to fund infrastructure-building projects and create a strategic plan to enable optimal use of big data, along with "a governance structure that aligns scientific leadership with resource management and oversight." There are two parts: "Big Data to Knowledge" (BD2K) to create a catalog of research datasets to facilitate data location, create standards and data-sharing policies, and "InfrastructurePlus" to help with large-scale computing and storage capabilities. To encourage and facilitate the use of biomedical big data, BD2K funds research to develop methods, software and other tools, training, and promulgates a data ecosystem. Ten Centers of Excellence were funded to this end, and will produce use-cases, analytics methodologies and techniques for maintaining data integrity and security (http://datascience.nih.gov/bd2k).

The National Centers for Biomedical Computing (NCBC) was a funding mechanism towards building a national computational infrastructure for biomedical computing. Seven centers were funded to develop software and other tools to create methods in modeling and simulation, and to develop means for data aggregation and sharing.

8.2 Public and Private Initiatives to Develop Large Cohorts

Studies linking genomics databases from biorepositories with patient medical records have been ongoing for some time (cf. Denny et al. 2010; Kho et al. 2012). In 2007, the NIH (through the National Human Genome Research Institute) created a nation-wide network of biorepositories and research centers to expand the capacity to do genome-wide association studies (GWAS) using information derived from electronic medical records. The Electronic Medical Records and Genomics network (eMERGE) became a large-scale collaboration to develop methods and procedures to do digital phenotyping of a variety of illnesses, then conduct GWAS studies on the data, as well as to develop procedures for wide-scale, secure data sharing (Gottesman et al. 2013). Consent processes were harmonized, using a "broad consent" model, but data access procedures had to be developed to enable the aggregation of data sets across sites and to allow access for researchers outside the network, when warranted (McGuire et al. 2011).

Beyond linkage of existing data, however, is the growing phenomenon of large-scale recruitment of individuals to volunteer personal health information and links to their medical records, sometimes with the addition of mobile health devices. For example, the Million Veterans Project, initiated by the Veterans Administration, already captures longitudinal data on all individuals who have done military service in the U.S. plus members of their families. The project additionally obtains survey data and trolls Facebook entries for behaviors and patterns. Predictive analytics can then be used to predict behavioral issues such as suicidal tendencies (Roski et al. 2014). Most recently, President Obama announced a new nation-wide Precision Medicine Initiative, giving it prominence as one of his four major goals for his administration. A Precision Medicine Initiative Working Group was created to facilitate the goals in the NRC report discussed at the beginning of this chapter: namely, the elimination of institutional, cultural and regulatory barriers to the free flow of data. The Working Group is currently trying to grow a National Research Cohort, similar to efforts in the U.K. and other countries, consisting of a million individuals from whom biospecimens and informational data, including data streamed from consumer and medical mobile health devices, can be collected longitudinally (Hogle forthcoming). Public workshops were held in the summer of 2015 to recruit disease advocacy groups, researchers, and individuals to participate. At the time of this writing, considerable promotion of the initiative continues. Additional public, private, and hybrid initiatives are proliferating, each aiming to restructure data relations among patients, investigators and institutions.

Such initiatives have opened up significant new market opportunities for entrepreneurial start-ups and large pharmaceutical and consulting firms alike. About $6.5 billion in new venture capital was invested in digital health in 2014, up from $1.2 billion in 2010, and many new businesses are arising from opportunities to broker data, sell smart phone apps and other software, capitalize on genomic

information capture, and otherwise use data as a product for health care, claiming to turn messy, unstructured information into an economic asset, or as Murdoch and Detsky (2013) put it, go "from refuse to riches" (See also Hogle forthcoming).

9 Discussion

Technologies and the practices surrounding them cannot be understood apart from the way they are embedded in socio-technical systems. In the case of big data and biomedicine, this includes the various devices and instruments involved (software, hardware, data storage, gene sequencing, mobile health devices) and the algorithms used and theories behind them, but also the users and decision-makers (clinicians, payers, public health or disease researchers, patients, policymakers, payers), the standards and norms, legal codes and ethical guidelines. In the U.S., the uptake of big data into biomedical research and clinical practice occurs within a particular historical moment of escalating health care costs and subsequent major reform efforts, but also ongoing disturbances about surveillance and aggressive data aggregators who see individuals' personal information as a valuable asset. Some of the specific struggles with setting health IT policy in the U.S. have to do with the particular ways in which health care has been traditionally organized and funded.

As I have shown, efforts to instantiate data-driven techniques into routine practice require major alterations of institutions, shifts in funding, and new laws and policies to promulgate their use while beginning to close off other alternatives. The social-technological systems being formed around data in contemporary biomedicine entail new economies but also social justice struggles, as traditional ways of thinking about protections and the proper flow of information are disrupted. There are resistances, political exigencies, and unanticipated outcomes that have thus far prevented biomedical big data governance and operating systems from becoming an established infrastructure. It is at these points of dissonance that infrastructures and the politics that surround them become visible. Exposing these dissonances and tensions provides an opportunity to examine ethical issues with a broader view.

A few points from the analysis I have presented bear highlighting. First, such a massive infrastructure-building project entails shifts in funding and priorities. Money is being diverted from coffers normally devoted to service payments for older and poorer Americans (Medicare and Medicaid) as well as funding previously used for health outcomes research to pay for the re-oriented programs to promote data-driven medicine. The most recent proposed Congressional budget would eliminate the Agency for Healthcare Research and Quality and the Patient-Centered Research Outcomes Institute (PCORI), created by the Affordable Care Act to identify treatments that work best.[37] Arguments to support the shift suggest that

[37] See http://www.hhs.gov/sites/default/files/budget/fy2016/fy-2016-budget-in-brief.pdf

big data is a more cost effective way of producing knowledge about outcomes than conventional outcomes research.[38]

The persistent call to include patient-generated data with structured, quantified measurements and the use of data-driven designs for discovery research and clinical trials changes the way knowledge is produced. Such measures alter conventional hierarchies of knowledge and go outside of institutions through which knowledge is usually produced. Notions of 'expertise' are challenged when 'non-medical' experts (data scientists, but also including patients themselves) participate more directly in processes of defining health, illness, and care. Critical data studies scholars point to epistemological issues inherent in data-driven approaches, including potential spurious results and dubious assumptions being used when analysis is driven by associative relations rather than seeking causative relations (Hoffman and Podgurski 2013; Kitchin 2014; Stevens forthcoming). Yet when big data methods are instantiated into funding mechanisms or incentives mandated through policy, and into regulatory requirements used to evaluate products for safety and efficacy, they will take precedence over other ways of knowing.

The "free the data" rhetorics are still very much in evidence; closer inspection of actions show the work involved in making data flow in the desired directions while making data practices a part of routine workflow (Roski et al. 2014).

Finally, observers should ask what values are being inscribed into biomedical big data infrastructures. Big data has been granted considerable perceived authority to solve problems in healthcare and biomedicine; at the same time, there is potential for tremendous impact on social and political life. There are implications for surveillance of citizens, but also social sorting; that is, the use of information to create profiles that may have consequences for the way individuals are viewed by payers, by consumer marketing groups, and others (Lyon 2003; Pasquale and Ragone 2014). Conventional concepts of autonomy are also challenged when data is collected ubiquitously and continually, and data is scooped from both "medical" and "consumer" domains to sketch portraits that may be used to characterize individuals. It is also important to ask who is *not* served by the flow of funds and information. Bayer and Galea question the push to precision medicine because of how it redirects public health expenditures and efforts toward relatively narrow sets of predominantly genetic conditions (2015). They raise the point that such massive programs do little to improve the nation's health in light of ongoing health disparities and population-level health needs. While enormous organizational efforts are preoccupied with interoperability of data sets, individuals are likely to worry more about the continuing difficulty in acquiring information that is most relevant to their own well-being, and how to prevent it from flowing in a way that may create harms for them.

[38] Additional criticisms of PCORI, however, suggest that funds have been used more for promotional and justificatory activities than meaningful outcomes research, and have been diverted to special interests, including industry lobbying organizations. It is sometimes difficult to tease out criticisms.

I have argued that attending to the less-visible issues of infrastructures shows how credibility and authority is established for novel technologies such as big data. At the same time, exposing some of the agreements, conflicts, and negotiations among participants shows how socio-technical systems arising in contemporary biomedicine are neither determined by particular technologies, nor by social conditions in which they arise. Rather, they are dynamic, contested, and interactive. The intense focus on privacy and data security in big data policy and ethical critiques to the exclusion of broader questions obscures some of the transformations I have described. When studying ethical and societal issues around technologies, it is important to expose the actual practices of how they come to count. Attending to the less-obvious issues of infrastructure-building enables an understanding of the changes taking place, and provides an opportunity to affect choices that may be consequential for health care systems and individuals.

Acknowledgement Research for this chapter was supported by the University of Wisconsin Graduate School Interdisciplinary Award. I am grateful to Joseph Wszalek for his contributions to research on relevant legislation.

References

Adler-Milstein, J. 2013. Operational health information exchanges show substantial growth but funding remains a concern. *Health Affairs* 32: 1486–1492.

Adler-Milstein, J., and A. Jha. 2013. Health care's "Big Data" challenge. *American Journal of Managed Care* 19(7): 537–538.

Agency for Healthcare Research and Quality. 2014. *A Robust Health Data Infrastructure*. Report No. 14-0041-EF prepared by JASON. McLean, VA.

Avorn, J., and S. Kesselheim. 2015. The 21st C Cures act: Will it take us back in time? *New England Journal of Medicine* 372: 2473–2475. doi:10.1056/NEJMp1506964.

Barocas, S., and H. Nissenbaum. 2014. Big data's end run around anonymity and informed consent. In *Privacy, big data and the public good: Frameworks for engagement*, ed. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, 44–75. New York: Cambridge University Press.

Bayer, R., and S. Galea. 2015. Public health in the precision-medicine era. *New England Journal of Medicine* 373(6): 499–501.

Blumenthal, D. 2010. Launching HITECH. *New England Journal of Medicine* 362(5): 382–385.

Brailer, D. 2009. Presidential leadership and health information technology. *Health Affairs* 28(2): w392–w398.

Charles, D., M. Gabriel, and M. Furukawa. 2014. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2013 Office of the National Coordinator for Health Information Technology, Data Brief No. 16. Available at http://www.healthit.gov/sites/default/files/oncdatabrief16.pdf

Cohen, G., R. Amarasingham, A. Shah, B. Xie, and B. Lo. 2014. The legal and ethical concerns that rise from using complex predictive analytics in health care. *Health Affairs* 33(7): 1139–1147.

Collins, F. 2004. The case for a US prospective cohort study of genes and environment. *Nature* 429: 475–477.

Denny, J., M. Ritchie, D. Crawford, et al. 2010. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 122: 2016–2021.

Edwards, P. 2003. Infrastructure and modernity: Force, time, and social organization. In *The history of sociotechnical systems. Modernity and technology*, ed. T.J. Misa, P. Brey, and A. Feenberg, 185–226. Cambridge, MA: The MIT Press.

Edwards, P., G. Bowker, S. Jackson, and R. Williams. 2009. Introduction: an agenda for infrastructure studies. *Journal of the Association for Information Systems* 10(5): 364–374.

Evans, B. 2009. Congress' new infrastructural model of medical privacy. *Notre Dame Law Review* 84(3): 586–654.

Executive Office of the President. 2014. Big data: Seizing opportunities, preserving values. Available online at https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Findlay, S. 2015. The FDA's sentinel initiative. Health Policy Briefs. Available at: http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=139

Fredrikson, M., E. Lantz, S. Jha, D. Page, and T. Ristenpart. 2014. Privacy in pharmacogenetics: And end-to-end case study of Warfarin dosing. Proceedings of the 23rd USENIX symposium. Available at https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf

Gottesman, O., H. Kuivaniemi, G. Tromp, W.A. Faucett, et al. 2013. Electronic medical records and genomics network: Past, present and future. *Genetics in Medicine* 15(10): 761–771.

Grossman, C., B. Powers, and J.M. McGinnis (eds.). 2011. *Digital infrastructure for the learning healthcare system*. Washington, DC: National Academies Press.

Gymrek, L., et al. 2013. Identifying personal genomes by surname inference. *Science* 339: 321–324.

Halamka, J. 2014. Early experiences with big data at an academic medical center. *Health Affairs* 33(7): 1132–1138.

Hoffman, Sharona, and Andy Podgurski. 2013. The use and misuse of biomedical data: Is bigger really better? *American Journal of Law & Medicine* 39: 497–546.

Hogle, L. forthcoming. Big data assemblages in healthcare. *BioSocieties 11(3).*

Hood, L., and S. Friend. 2011. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology* 8(3): 184–187.

Jasanoff, S. 2005. *Designs on nature: Science and democracy in Europe and the United States*. Princeton NJ: Princeton University Press.

Kahn, Howie. 2013. Who *really* owns your personal data? http://www.details.com/culture-trends/critical-eye/201305/sharing-biodata-on-apps-and-devices. Accessed 25 Aug 2013.

Kaye, J. 2015. The tension between data sharing and the protection of privacy in genomics research. In *Ethics, law and governance of biobanking*, ed. D. Mascalzoni. New York: Springer.

Kho, A.N., M.G. Hayes, L. Rasmussen-Torvik, et al. 2012. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association* 19: 212–218.

Kitchin, R. 2014. Big data new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12. doi:10.1177/2053951714528481. Accessed 20 June 2015.

Kohane, I. 2011. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics* 12: 417–428.

Law, J., and W. Bijker (eds.). 1992. *Shaping technology/building society: Studies in sociotechnical change*. Cambridge, MA: MIT Press.

Loukides, G., J. Denny, and B. Malin. 2010. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association* 17: 322–327.

Lupton, D. 2014. The commodification of patient opinion: The digital patient experience economy in the age of big data. *Sociology of Health and Illness* 36(6): 856–869.

Lyon, D. 2003. *Surveillance as social sorting: Privacy, risk and digital discrimination*. New York: Routledge.

Malin, B., D. Karp, and R. Scheuermann. 2010. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigational Medicine* 58(1): 11–18.

McGuire, A.L., M. Basford, L.G. Dressler, et al. 2011. Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience. *Genome Research* 21: 1001–1007.

Mittelstadt, B.D., and L. Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341. doi:10.1007/s11948-015-9652-2.

Mittelstadt, B., N.B. Fairweather, M. Shaw, and N. McBride. 2014. The ethical implications of personal health monitoring. *International Journal of Technoethics* 5(2): 37–60.

Murdoch, T., and A. Detsky. 2013. The inevitable application of big data to health care. *Journal of the American Medical Association* 309(13): 1351–1352.

National Research Council Committee on a Framework for Developing a New Taxonomy of Disease. 2011. *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. Washington, DC: National Academies Press.

Neff, G. 2013. Why big data won't cure us. *Big Data* 1(3): 117–123.

November, J. 2012. *Biomedical computing: Digitizing life in the United States*. Baltimore: Johns Hopkins University Press.

Office of the National Coordinator for Health IT. 2014. *Report to congress: Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information*. Washington, DC. Available at https://www.healthit.gov/sites/default/files/rtc_adoption_and_exchange9302014.pdf.

Office of the National Coordinator for Health IT. 2015a. Connecting health and care for the nation: A shared interoperability roadmap. Available online at http://www.healthit.gov/sites/default/files/nationwide-interoperability-roadmap-draft-version-1.0.pdf

Office of the National Coordinator for Health IT. 2015b. Federal Health IT Strategic Plan 2015–2020. Available online at http://www.healthit.gov/sites/default/files/federal-healthIT-strategic-plan-2014.pdf

Pasquale, F., and T. Ragone. 2014. Protecting health privacy in an era of big data processing and cloud computing. *Stanford Technology Law Review* 17: 595–653.

President's Council of Advisors on Science and Technology (PCAST). 2014. Big data and privacy: A technological perspective. Report to the President (May). Available online at https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Ranck, J. 2014. How connected health, public-private cooperation, and big data can revolutionize health care. *Forbes*. http://www.forbes.com/sites/benkerschberg/2012/07/06/how-connected-health-public-private-cooperation-and-big-data-can-revolutionize-health-care/. Accessed 25 July 2014.

Redhead, S. 2009. The Health Information Technology and Economic and Clinical Health (HITECH Act). Congressional Research Service, Report No. R40101.

Ricciardi, L., et al. 2013. A national action plan to support consumer engagement via e-health. *Health Affairs* 32(2): 376–384.

Roski, J., G.W. Bo-Linn, and T.A. Andrews. 2014. Creating value in health care through big data: Opportunities and policy implications. *Health Affairs* 33(7): 1115–1122.

Saha, K., and L.F. Hogle. 2014. Allying with donors to link health and medical information with stem cell lines can advance disease modeling while enhancing data access. *Cell Stem Cell* 14(1): 559–560.

Schadt, E. 2012. The changing privacy landscape in the era of big data. *Molecular Systems Biology* 8(1): 612–617.

Singer, N. 2014. When a health plan knows how you shop. *New York Times*, June 28. Available at http://www.nytimes.com/2014/06/29/technology/when-a-health-plan-knows-how-you-shop.html

Star, S.L., and K. Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111–134.

Stevens, H. 2013. *Life out of sequence: A data-driven history of bioinformatics*. Chicago: University of Chicago Press.

Stevens, H. forthcoming. Hadooping the genome: The impact of big data tools on biology. *BioSocieties* 11(3).

Sulzicki, M., D. Atkins, C. Schilling, et al. 2012. A model to predict risk of non-adherence to medications highlighted in CMS star-ratings. *Value in Health* 15: A164 (PRM30).

Terry, N. 2012. Protecting patient privacy in the age of big data. Indiana University Legal Studies Research Paper 2013–04.

Thune, J., L. Alexander, P. Roberts, R. Burr, and M. Enzi. 2015. Where is HITECH's $35 billion dollar investment going? *Health Affairs Blog*, March 4. Available at: http://healthaffairs.org/blog/2015/03/04/where-is-hitechs-35-billion-dollar-investment-going/. Accessed 20 July 2015.

Vaidya, M. 2014. As funds to sync health records dwindle, research could suffer. *Nature Biotechnology* 20(11): 1225–1226.

Weiss, J., S. Natarajan, P. Paissig, C. McCarthy, and D. Page. 2012. Machine learning for personalized medicine: Predicting primary MI from electronic medical records. *AI Magazine* 33(4): 33–45.

Xu, H., et al. 2011. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annual Symposium Proceedings* 2011: 1564–1572.

Ethical Reuse of Data from Health Care: Data, Persons and Interests

Peter Mills

Abstract Advances at the intersection of information technologies, data science, biomedical research and health care, have discomfiting implications for reliance on the conventional tools of data protection and information governance. A 2015 report from the Nuffield Council on Bioethics proposes an approach to the design and governance of data initiatives that is both dynamic and cooperative in order to address the fluctuating interests engaged by uses of data in which there is a public interest. This chapter develops elements of that approach to argue that organising data initiatives as social practices that respect certain principles can help to establish and meet morally reasonable expectations about data use, by grounding them in a dynamic relationship between social norms, individual freedoms and professional duties.

1 Introduction: The Nuffield Council Report

The Nuffield Council on Bioethics is an independent body, established in 1991, that examines and reports on ethical issues relating to advances in biological and medical research. It is funded jointly by the Nuffield Foundation, the UK Medical Research Council and the Wellcome Trust. In the 2013 the Council undertook to examine bioethical issues raised by developments in data science and information technologies. This resulted in a report, published in 2015, entitled *The collection, linking and use of data in biomedical research and health care: ethical issues* (Nuffield Council on Bioethics 2015). In the present chapter I will foreground one line of analysis suggested by the report, concerning the use of data in health care systems, and outline the main features of the approach recommended in the report. The intention will be to characterise a difficulty for health care information governance that arises from the adoption of 'Big Data-like' approaches to data use, and to suggest why the Council's recommended response is particularly appropriate

P. Mills (✉)
Nuffield Council on Bioethics, 28 Bedford Square, London, WC1B 3JS, UK
e-mail: pmills@nuffieldbioethics.org

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_18

to cases, such as health care, in which issues of public interest are at stake. In doing so it will be illuminating to say something about the development as well as the content of the Nuffield report. I will begin, though, by describing the Council's conceptualisation of the novel context created by advances at the intersection of information technologies, data science, biomedical research and health care, and the discomfiting implications that these have for reliance on the conventional tools of data protection and information governance.

2 Data Fetishism

Perhaps the report's first conclusion was that (*pace* the editors of this volume) it is increasingly difficult to speak of 'biomedical data' and quite possibly mistaken or, at least, misleading to do so. In any meaningful occurrence data are invariably plural. A single datum functions, perhaps, only like the Cartesian *cogito*: a signal that there is something rather than nothing. Data as they appear are always composed and configured in sets (and, therefore, potentially open to decomposing and recomposing in different ways). Understanding both 'data quality' – in terms of the limitations imposed by the conditions and methods of data generation – and the way in which data sets are structured by intentionality and purpose (their 'aboutness'), is therefore a key problem when it comes to assigning significance to them. It is perhaps trite to observe that the informative value of data is highly dependent on the context in which they are placed, although consequences follow from this that the Nuffield report argues have not been taken seriously enough in information governance practice. These consequences become particularly important when intensified by the power and promiscuity of 'Big Data'.

2.1 The Data Protection Paradigm

What we may call the 'data protection paradigm' is a particular way of managing potentially competing human interests in the use (or 'processing') of data. It is orchestrated by a number of key concepts, such as 'personal data' (and its implicit negative image, 'anonymous' data), 'sensitive data' and even 'health-related data'. Applying these concepts to actual data is problematic because the operation is highly dependent on the intentional context and epistemic circumstances in which it is carried out. Let us call this the problem of schematism. The properties of being 'personal' or 'anonymous', or of being 'health-related', for example, depend substantially on the possession of data and access to data enjoyed by the agent performing the schematism (the 'data processor') and on the purpose for which they are processing the data (which may be closely related). One has to situate and orientate the data processor in a data environment, and the semantic properties any given set of data have for them at any given time depend on this disposition.

That the schematism becomes problematic is not so much a failing of the conceptual scheme of data protection itself but of the way in which it is put to use in the work of information governance. This can happen when the concepts ('sensitive', 'health-related', etc.) become hypostatised – treated as if they were properties of the data themselves – when data are transferred between significantly different epistemic and intentional contexts (or, which amounts to the same thing, the epistemic context changes owing, for example, to the availability of new data). In biomedical research and health care practice we see this, for example, in simple standards of anonymisation or pseudonymisation (removing name, address and date of birth or replacing these with a unique code) or with the assumption that data collected in a health record have equal and distinct sensitivity (being routinely more sensitive than other data). This hypostatisation may underwrite certain norms, behaviours and 'rules of thumb'. One such is the injunction, hitherto common in the English NHS, to 'consent or anonymise', that is, to process data outside the context of collection only when it has been rendered 'anonymous' or with the consent of the data 'subject(s)'.

Having rules like 'consent or anonymise' is, of course, pragmatic and, in this sense, reasonable. If it is sometimes misleading to speak of data as 'identifiable' or 'anonymous' it is easier to do so than to surmount the cognitive challenge involved in a full consideration of their capacity to identify, and even more so than to consider the ramifications of identification for the privacy interests of those to whom the data relate, every time they are 'processed'.[1] When equivocations arise, or when mistakes or breaches occur as a result of everyday data handling practices, they may be hived off and dealt with (often in relation to such a pragmatic standard of reasonableness).[2] Data protection law and practice are hedged about in such a way. One reason such practices have held up reasonably well may be that the epistemic and intentional context of data processing is often fairly close to that of data collection and can often be reasonably well characterised or anticipated. Recent developments in the use and exploitation of data, however, make this incresingly insecure.

2.2 Big Data

The reason that the Nuffield Council set itself to consider the ethics of data use in biomedical research and health care can be explained as the conjunction of a number

[1]In the interests of economy these arguments, which are elaborated elsewhere (e.g. Nuffield Council on Bioethics 2015), are not developed or explored here. It will be assumed to be common ground that the concepts of anonymisation and consent are problematic for Big Data, in that it is implausible to suggest that anonymisation can be robust without further qualification and a mistake to believe that consent guarantees autonomy or is either necessary or sufficient to secure the protection of all relevant personal interests.

[2]It is probably worth remarking that, nevertheless, perhaps the biggest sources of negative impact from misuse of data remain errors of competence and maladministration (Laurie et al. 2014).

of developments, all of which will be reasonably familiar. The first is the escalating accumulation of data in general and in health care and biomedical research in particular. In the UK, especially, the accumulation of data in National Health Service (NHS) records and information systems is internationally unparalleled. This accumulation of data was greatly facilitated by the adoption of computerisation and the fact that professional interactions, like a great many social interactions, are electronically mediated, each one generating an electronic trace. In many cases these will be patient records and transactional data recorded for clinical or administrative purposes, but they are abundantly supplemented by quantities of system-generated metadata. While the NHS benefitted from its national organisation and a degree of standardisation (following the introduction of the Lloyd George record in 1911) a limiting factor, in the case of the NHS, has been the curiously late and haphazard adoption of health service information technology (Wanless 2002 and Wanless et al. 2007), which a number of overambitious, ill-judged or poorly executed initiatives, such as the expensively abandoned National Programme for IT, have attempted to overcome.

As well as an increase in routine clinical and administrative data collection facilitated by the widespread adoption of information technology, significant new sources of data have come on stream. Of particular interest in the health context are genome sequencing, other 'omics', and medical imaging (functional magnetic resonance imaging of the brain, for example). Large scale genomic data generation is the subject of another recent UK government initiative, the '100,000 genomes' project (Genomics England Ltd 2015). The size of the data sets generated by these new technologies present technical challenges of interpretation and knowledge discovery: processing genomic data, for example, involves mathematical procedures to clean data, impute the value of missing data and represent it in meaningful form. These need to be controlled carefully. An infamous imaging experiment in which positive 'brain activity' was detected in a dead Atlantic salmon as a result of a standard investigative methodology vividly evokes this point (Bennett et al. 2009). Developments in bioinformatics and data science more generally have responded to this; the point of drawing attention to it here is to underline the difference between the meaningful health or biological information and the essentially metaphysical notion of 'raw' data.

The discourses of data protection and Big Data speak of related problematics, one 'extensive' (the containment of data within or transfer of data between epistemic contexts in which they acquire a distinctive significance) and the other 'intensive' (the representation of internal relations within data sets at more abstractly structured orders of representation to produce information). In this sense Big Data aggravates the schematism problem (where it problematises the context) or even inverts it: it is a claim that is often made for Big Data, though one equally often disputed, that sufficiently rich data may themselves generate a novel level of meaning through a kind of 'semantic autopoiesis', generating 'hypothesis-free' knowledge (Mayer-Schönberger and Cukier 2013).

3 Data and Persons

When we look back from the perspective of Big Data, we find that the humanistic categories of 'subject', 'purpose', 'controller' and their relations have slipped out of sight: intentionality has become methodology, and purpose has become abstract insight and value generation. Where data protection strives to concretise data according to its conceptual scheme, Big Data tends to etherialise data or, at least, to liberate them from the epistemic and purposive context of collection/generation in order to re-territorialise them elsewhere. This is not simply a matter of framing or perspective, but also one of ethics and of politics to the extent that Big Data appears to exert a force that continually undermines the data protection paradigm. The following sections argue that the opportunities suggested by Big Data create a pressure to realise expectations that puts further stress on the categories of information governance, in particular, exposing the conceptual distance between the 'personal' and the 'private'. They support an argument for moving away from reliance on the nostrums of information governance in favour of a continual rethinking of the relationship between public and private interests in data use.

3.1 Double Bind

The focus of the Nuffield report is not so much on the question of scale *per se* or on the question of mining large data sets to extract novel insights that were not discoverable at a smaller scale, but more on the question of repurposing and linking data, and translating them into new semantic contexts (that is, with what were described above as 'extensive' questions). The polymorphous potential of a Big Data, has given rise to a significant new attitude towards data generated in health care that sees them as a valuable resource that may be reused indefinitely, linked, combined or analysed together with data from different sources, to generate new insights and knowledge. This means, at a humanistic level, often handing them over to other actors, with their own aims and interests. The UK government was not chary of seizing on the promise of Big Data to promote the UK as an incubator of scientific advance and economic value (Department for Business, Innovation and Skills 2011; Willetts 2013). The Nuffield report, therefore, draws attention both to the developments in knowledge and resources, and to the scientific, political, and economic drivers to exploit them. These result in two imperatives:

- to generate, use and extend access to data (because doing so is expected to advance research and make public services more efficient); and, at the same time,
- to protect privacy as a requirement of human rights law (and the more access to data is extended, the greater the risks of abuse).

Though not logically contradictory, these injunctions may become practically irreconcilable when certain features and consequences of contemporary

technological and epistemic contexts are taken into account. Among the most often discussed consequences are those that bear on the commonplaces of information governance: that the anonymity of the subjects of individual-level data is (or will become) effectively impossible to maintain and that the value of consent in governance (insofar as it is thought to express an 'informed' choice that can meaningfully be withdrawn) becomes significantly attenuated.[3] To the extent that the mechanisms that are still relied on to protect individual privacy in the data protection paradigm are exposed to failure by the very techniques that are supposed to enable the exploitation of data for public good, the twin imperatives result in something like a 'double bind' (Bateson 1972).

3.2 Not All 'Personal' Data Are 'Private'

The data protection paradigm hangs on data that are in some way related to individual people and that are also *thought* as being about those people: 'personal data'. Not all personal data, however, will be *private*. Of course, like the property of being 'personal', the property of being 'private' is not one that belongs to data at all, or does so only contingently. Privacy refers to a certain kind of relationship between moral agents, which is expressed in various material and informational contexts (private property, withdrawal from observation by others, inviolability of correspondence, etc.).

The unifying notion of privacy as "a relevant state of separation defined and mediated by particular standards" can be expressed through of 'norms of exclusivity' that regulate access to individuals or groups (Taylor 2012). We recognise that the enactment of this separation (and its regulation) has an important function in establishing and maintaining the structure of social and interpersonal relationships: acting as a token of friendship, cementing social bonds, promoting trust, and encouraging reciprocal sharing. In relation to information, privacy is usually understood to be about access to and disclosure of data. This, however, leaves unexplicated the important role played by the understanding of the semantic context.[4] What is schematised as 'private' can differ and alter depending on social norms, the specific context, and the nature of the relationship between the individuals (or groups) and others (at the limit, *all* others) concerned.

[3]These were, respectively, propositions 15 and 16 in the Nuffield report (Nuffield Council on Bioethics 2015).

[4]Manson and O'Neill make the distinction between *information* (thought as content) and *informing*, the effect of which relies substantially on a background of knowledge, competencies and dispositions (Manson and O'Neill 2007).

3.3 Some 'Non-personal' Data Are Privacy-Affecting

Just as not all 'personal' data are 'private' (or privacy affecting), so data that are not 'about' a particular – or even any – individual may have an impact on individuals' privacy. Perhaps the most insidious privacy-affecting impacts are those that data dependency in professional, public and social transactions have on norms of disclosure, often through making disclosures a condition of access to services or preferentially affecting their cost. A salient example is the bargain struck by users of social media that involves their data being made available for marketing and other purposes (encapsulated in the meme 'when something online is free, you're not the customer, you're the product'). But while 'opting in' to services that have become part of the warp and weft of modern social and professional life involves complex trades off, 'opting out' of services to which there is arguably a moral entitlement makes this a matter not of the fairness of the trade, but about the fairness of trading.

There is a premonitory postscript to the hackneyed 'big data bogeyman' story – the one in which the US retailer, Target, identified a pregnant teenage girl from her purchasing behaviour and inadvertently revealed her condition to her successively irate and chastened father (see, for example, Mayer-Schönberger and Cukier 2013). In this subsequent case an academic sociologist recounts the difficulties and disadvantages of her attempt to mask her own pregnancy from big data analysis, concluding that this both requires a high level of technical knowledge and entails significant social, financial and (potentially) legal costs (Vertesi 2014). We may take this as a premonition of the way in which data dependency that comes about largely through the free market can have consequences that are arguably contrary to both privacy and public interest. The significance of this premonition becomes clearer when a well-established social good is at stake, where it is subject to transformative technological innovation and, especially, where market mechanisms are allowed to operate freely.

Perhaps the preeminent case of a social good from which opting out may entail serious personal consequences is health care. Extending the use of data collected in health care settings can therefore have morally significant implications, since the consequences of withdrawing from it (or of being prevented from doing so), both for the individual and for the population, are potentially grave. Innovations in technology and data use (whether clinical, scientific or administrative), how options are bundled together (particularly with digital technologies) and whether patients should be expected to opt in or opt out of them, cannot be matters of indifference.

3.4 Some 'Personal' Data Have Public Implications

One of the main reasons to be concerned about the wider use of data from healthcare and biomedical research is their bearing on the public interest. Data collected in health services and biological research, in particular, are of value not only in relation

to the treatment of the person from whom they are collected but also in that they contribute to scientific understanding, research and development, and provide an evidence base for policy or administrative decision making. Similarly, failure to use data in the public interest affects private interests because, as members of the public, we are severally (in the case of a medical advance and, in particular, in the case of 'personalised' medicine) or collectively (in the case of a public health measure) beneficiaries.

The Nuffield report draws attention to two questions about public interest: how to determine the *content* of the public interest with regard to the use of data and the *force* that should be accorded to it, particularly when there are competing individual interests at stake. In the case of data initiatives in health care addressing this takes the form of a critical reflection on the conditions of the tension between individual interests and the 'other' of the public interest, namely an examination of the way in which individual and public interests are mutually implied.

> Consequently, our problem is not finding a 'balance' between privacy and public interest for a data initiative, but resolving a double articulation, between the private interest in protecting privacy and promoting the public good, and the public interest in protecting privacy and promoting the public good. We all have interests on both sides, private and public, as individuals, members of families, groups, communities and nations. Navigating among these different relationships with other individuals, professionals and institutions requires a subtle negotiation of many different norms of information access and disclosure, of when and how they may be modified and where hard and fast limits should be drawn. (Nuffield Council on Bioethics 2015)

This reflection on the mutual implication of privacy and public interest, on privacy as concerning norms and morally reasonable individual preferences (that must be morally reasonable when assessed in relation to those norms), and public interest as the assertion of morally reasonable norms and collective preferences, leads to the formulation of a question with regard to the moral basis of data initiatives in general, one that recasts the double bind – which seemed intractable so long as it embodied a contradiction between imperatives transfixing an actor in the data protection paradigm – as a co-operative problem, namely: *how may we define a set of morally reasonable expectations about how data will be used in a data initiative, giving proper attention to the morally relevant interests at stake?*[5]

[5]Finding a use of data that is mutually acceptable to those whose morally relevant interests are at stake gives rise to the requirement that the expectations about the use of data form a coherent 'set', one that does not contain contradictory elements (although surpassed contradictions may be preserved at levels below that of the collective agreement). Furthermore, respect for those whose interests are at stake (which, in some cases are protected by rights and entitlements) means that it must be one that is capable of being articulated ('publicly statable') in a way that is meaningful, and understandable to those whose interests are at stake, such that an account of decisions can be given to them that they would recognise as reasonable (Daniels and Sabin 1997; Habermas 1990).

4 Persons and Interests

The approach that the Nuffield report took to addressing this question was to treat data initiatives as cooperative social practices or, rather, to take seriously the implications of data sharing being a social practice.[6] This means that the report does not start by asking what kinds of data are in play but rather by asking on what norms of privacy and disclosure particular uses of data trespass, and what interests they engage.

In response to the practical question of *how to define* a set of morally reasonable expectations the report recognises two main senses in which expectations may be 'morally reasonable'. The first is as conforming to some independent standard of moral reasonableness; the second is as the outcome of a process of moral reasoning.[7] Rather than offering a purely formal or procedural response, the Nuffield report made a set of recommendations that embodied elements of both of these approaches, relying on procedural elements (to ensure a situated outcome capable of accounting fairly for the interests actually at stake) within parameters set by standards of respect for persons and human rights (to avoid the possibility of morally perverse outcomes resulting from a purely procedural approach).

A crucial element of the procedural aspect is participation in the establishment of the set of expectations by 'those with morally relevant interests' in order to put into play both the expectations and the values associated with them. These agents should include a full range of interests: not only those who will handle or use the data, or the outputs of the initiative – the professionals involved – but also those who may be affected by the initiative, who may stand to benefit or be harmed by different outcomes, in particular, where the data may be related to them as individuals or groups.[8] The range of such interests and the form in which they precipitate into distinctive moral agents (individuals, groups, collectives) will differ from initiative to initiative, which is why there can be no complete universal set of norms, no 'one-size-fits-all', as the Nuffield report says. This means that any set of minimal or general norms, for example those encoded in high level legislation, is unlikely, on its own, to provide an adequate basis for governance: they may be a poor 'fit' for

[6]The Nuffield report remarks, as others have done, on the imprecise and loaded use of language in the relevant literature. 'Data sharing', with its connotations of beneficence and mutuality, is an example of this although the term seems appropriate in this context where the norms of exclusivity are established through the involvement of relevant interests to which it restores an appropriate level of moral agency.

[7]Of course, second order questions arise in each case (for example, in the first case, about the origin or justification of the 'independent' standard; in the second, regarding the fairness of the process). The report does not try to resolve these but it keeps them in play. Consequently the process is reflexive and interminable: its conclusions are provisional and persuasive rather than decisive, and the standard is treated as conventional rather than categorical.

[8]This is not only those *from whom* the data may have been collected but also those *to whom* people the findings might be applied. If the findings inform decisions relating to public policy, this entails a case for the engagement of the 'public' in general (or their representatives).

particular data initiatives, both in the sense of being unnecessarily restrictive and in the sense of being insufficiently sensitive to the interests involved.

It is through participation that the relevant norms and interests – and the nature of their relevance – are discovered. In the case of National Health Service patient data, for example, the scope of morally relevant interests will include the interests of those treated by and working in the NHS, as well as other potential beneficiaries such as health policy makers, academic and professional researchers. Through this process, without denying the existence of conventional rights or duties, what is brought into question are the circumstances or conditions in which the entitlements would be claimed or set aside, along with the relevant social norms. But this is not all: it also, importantly, provides a motor for the transformation of social norms and value commitments as they encounter each other in a process of respectful deliberation (Parker 2007).[9] This means that moral reasoning need not stop at the threshold of acceptability – of determining merely what is acceptable – but should have a continuing and constructive function, moving beyond what is merely tolerable to what is morally preferable for those whose interests are interdependent. This is especially appropriate where collective choices have opportunity costs, where the choice of a suboptimal option may foreclose a better one (Nuffield Council on Bioethics 2012).

4.1 Establishing Norms, Freedoms and Duties

Three key elements within the set of expectations are the norms of privacy and disclosure that are at stake in any given data initiative, the form in which respect for moral agents is expressed, and the requirements of governance in place. These relate, first, to the common social norms, second, to the freedom of individual moral agents, and, third, to the duties of professionals, linking the spheres of, respectively, social relationships, private conduct and professional practice.

To elaborate this scheme I should like briefly to recall the genealogy of this tripartite scheme in the deliberations of the Nuffield Council working party that steered the development of the report. The scheme grew out of an attempt to categorise the possible 'moral bases' for data sharing and the conditions in which each would apply. The idea was that these categorical norms would provide a stabilising and justifying function for claims made about the adequacy and proportionality of the other measures. The scheme of moral bases originally comprised duty, solidarity, autonomy, reciprocity, and authority. These were not meant to describe the implied motivations of all the actors involved in a given data initiative (which are, by

[9] In this respect it has advantages over aggregative approaches that simply gather preferences as a basis for decision making (e.g. voting) or are simply designed as a take-it-or-leave-it 'offer' to optimise the number or type of those consenting to participate. Nevertheless, owing to the fact that it is not practical to involve *all* those who have an interest, it is still necessarily only indicative (rather than representative).

hypothesis, confused and inconsistent) but rather the character of relations that would legitimately prevail at the 'public' level. Thus there would be cases in which it was morally reasonable to *require* the sharing of data, for example where public health was at stake, as in the case of an epidemic.[10] Hence, while respectful of individuals, the expression of individual agency in the form 'opting out' of data sharing in such a case would be morally impermissible.

The hypothesis was that the character of this 'moral basis' for data sharing would correlate with the way in which respect is shown for moral agents involved in the data initiative *because of* the way in which individual and public interests are mutually implied. As we proceed through the classification, different moral bases, by hypothesis, will underwrite different ways of expressing respect for individual preferences. One significant way in which this is done is through the rather vexed gamut of practices described as consent, although often less in relation to the giving of choice and more in relation to the choice given. Thus, in the hypothetical scheme, the different moral bases would stand as justification for different expectations about consent, ranging from none, where compliance is a universal duty (or a duty for all members of a definable class of persons), through implicit consent where there is a subsisting social or democratic licence, through general and more specific forms where there are particular requirements because important freedoms are recognised as being at stake. For example, where the moral basis for data sharing is solidarity (where there may be a prior acceptance that individuals should be willing to carry costs on behalf of others) explicit or detailed consent of individuals may not be required, although individuals may be free to opt out and should not be coerced to participate (Prainsack and Buyx 2013).[11] More specific forms of consent, on the other hand, become important where the transaction has a more economic character (where it is a matter of equitable trading off of benefits and costs).

The third element, governance, concerns the assurance of moral conduct of others: individuals are entitled to have expectations of others using data, particularly professionals involved in data initiatives. These include expectations about *who* these others will be, and how their conduct will be governed. There is a public interest in ensuring that those involved in data initiatives discharge a moral duty of care owed to others, especially those whose interests place them in positions of vulnerability, a duty that is not exhausted simply by complying with subjects' consent. As with consent, different forms of governance will be appropriate to different relationships between the public and private interests involved.

In the original approach, norms of privacy and disclosure were foundational. It was part of the reflection on the various successes and failures of previous data

[10]Interestingly, this case has been made, although not entirely persuasively, in relation to the sequestration of health data for Big Data analysis supporting biomedical research, health service planning and delivery (e.g. predictive interventions) and policy. The case is broadly that, without the gains in efficiency promised by better data use, the National Health Service in England will become unaffordable and cease to exist, imperilling public health.

[11]Prainsack and Buyx offer a descriptive ethics of solidarity in *Solidarity: reflections on an emerging concept in bioethics* (Prainsack and Buyx 2011).

initiatives that the degree of attention paid to the underlying norms and the way in which consent and governance approaches were designed in accordance with these had been crucial to the viability of the initiative.[12] However, one of the further lessons from the consideration of particular data initiatives was that norms are historically and culturally – and technologically – specific. This means that as health data services develop and integrate with research programmes to form 'learning health systems' or hybrid 'health care and research systems' interests are reconfigured in response to these new profiles of opportunity and threat (Faden et al. 2013). Given the impact of technology on social norms and their possible transformation through deliberation, the rather fixed scheme of moral bases was therefore abandoned in favour of a more free floating relationship between norms, freedoms and governance, with norms thrown into the mix. Thus, the constrained deliberative element (constrained, that is, by principles of respect for persons and established human rights) was intended, independently of any schematic characterisation, to arrive at a configuration of norms, freedoms and governance that could not be regarded as necessary or pre-existing – could not be determined through research and analysis – but had to be established by co-production through the engagement of interests at stake.

4.2 Respect for Persons in Practice

The Nuffield report formulates four principles for respectful use of data in biomedical research and health care. Alongside the substantive principles of respect for persons and human rights, which provide the consistent 'guard rails' against morally perverse outcomes, the Council framed a principle of participation, that "the set of expectations about how data will be used (or re-used) in a data initiative, and the appropriate measures and procedures for ensuring that those expectations are met, should be determined with the participation of people with morally relevant interests". It further stipulated that this participation "should involve giving and receiving public account of the reasons for establishing, conducting and participating in the initiative in a form that is accepted as reasonable by all" (Nuffield Council on Bioethics 2015). Nevertheless, except with ideally closed data initiatives, not all bearers of relevant interests will want or be able to be involved and, in most cases, the best that can be achieved is 'fair representation' of the full range of values and interests at stake. This consideration foregrounds a set of second order questions (Who determines what interests are relevant? Who determines what is 'fair' representation?); furthermore the scope and internal constitution of 'morally relevant interests' will often fluctuate over time. Since these must remain problematic, the outcome of any deliberative process can only be provisional.

[12]In preparing the report we proposed and 'empirical typology' of data initiatives as a basis for reflection.

Hence the need for a further principle, that of 'accounting for decisions', that has a regulatory function. This faces in two directions: in the direction of social accountability to provide a mechanism for interests that did not or could not participate in the establishment of a data initiative to come to bear, and formal accountability through structures of legitimate judicial and political authority that can enforce the substantive principles of respect for persons and human rights, which provides protection against inequity and discrimination.[13] The principle of accounting for decisions therefore entails both 'giving an account' and 'being held to account', both formally (to provide for appeal for those who feel they have been unfairly marginalised) and in relation to the provisional 'social licence' for data initiatives.

5 Data Initiatives from the Perspective of Big Data

The rather more fluid and continuous encounter between interests recommended in the Nuffield report also responds to what is a potential difficulty for approaches that appear to rely on any definition of data context, albeit one for which the reference is shifted from the context of data collection to that of data use. In the Nuffield report the use context is expressed through the concept of a 'data initiative', defined as a purposive activity involving either or both of the following practices:

- data collected or produced in one context or for one purpose are re-used in another context or for another purpose ('repurposing') or for no well-defined purpose (data prospecting).
- data from one source are linked with data from a different source (or many different sources) to provide insight that was not available before the data were so linked.

The circumscription of such a context is important for all three of the elements of the set of expectations: to define the community of morally relevant interests and the corresponding norms, to make choices meaningful (where choice is an issue) and to determine the scope and measures of applicable governance. This is problematic, however, because the polymorphous promiscuity of Big Data is in many ways antithetical to such a movement of circumscription: the set of expectations will be continually undermined as new possibilities for extracting value from data arise. If we allow that Big Data expresses an idea that is less about the abundance of resources or the power of technology than a disposition towards the use of data, we

[13] "A data initiative should be subject to effective systems of governance and accountability that are themselves morally justified. This should include both structures of accountability that invoke legitimate judicial and political authority, and social accountability arising from engagement of people in a society. Maintaining effective accountability must include effective measures for communicating expectations and failures of governance, execution and control to people affected and to the society more widely." (Nuffield Council on Bioethics 2015)

can see how its impact in the field of health care through the leverage of 'health' data by information technology and data science must be to intensify the interpenetration and amalgamation of organisational, functional and disciplinary systems. This will be the case especially where feedbacks operate between interrogation and results – where the results may refine, or even define, the questions posed. The definition of a learning health system – or health and care system, or care and research system, or care and research and welfare system – which may reach beyond surgeries and hospitals, out into communities, whose nervous system is the internet and whose organs are sensors, interfaces and monitoring devices, may be complex, fluctuating and vague.

The consequences for information governance are that circumscription of data initiatives not only in terms of ontology (no longer 'health' data, merely data), but also in terms of location (they are virtually present everywhere), purpose (they are polymorphous, promiscuous) or property (they are disposed by interests not ownership) can only be partial and provisional. There is therefore an advantage in a dynamic approach that, by giving effect to the principles of 'participation' and 'accounting for decisions', allows the use of data to be continually referenced to the question of reasonable expectations in a changing context. (Whereas dynamic consent models have been proposed (e.g. Kaye et al. 2014) their effect is intended to be to maximise opportunities for the expression of personal choice rather than to optimise system design or participation. Insofar as they may operate like price signals in a marketplace, representing a set of uncoordinated choices, their 'invisible hand' may steer system development over several iterations but their tendency is to identify a common denominator among preferences rather than the optimum distribution of choices that is acceptable to any arbitrarily defined group (Taylor and Taylor 2014)[14]. As a collective activity, the participatory and accountable approach aims not at controlling the exchange of data but at establishing the underlying relationships that make different uses of data acceptable. It is perhaps a measure of the importance of these relationships, and the insufficient attention that they have

[14]There is not sufficient space here to engage fully with the question of dynamic consent. It should be sufficient, however, to observe that dynamic consent is not a new form of consent: consent has always been 'dynamic', in the sense of continuously subsisting, capable of withdrawal or subject to the imposition of conditions, etc. The introduction of electronically supported consent portals merely add a facility of communication to it that, granted, offer a number of advantages (ease of use, communication of results, etc.). They do not change the fundamental choices available: what they do is 'unstick' the inertia of 'up front' consent recorded on paper consent forms and undermine the rationale for the 'compromise' (if compromise it really is) of 'broad' consent (although dynamic consent can also be 'broad': again, it adds nothing new). The important question of policy is not about the choice to exercise freedoms but the determination of what freedoms may be exercised. The same point applies, *mutatis mutandis*, to solidarity: the voluntary expression of a more social disposition, a 'willingness to carry costs on behalf of others' (Buyx and Prainsack 2011), while it may be something to be encouraged, does not overturn the individualistic model of consent. A solidarity-*based* system is one in which the distribution of *options* to bear costs and enjoy benefits is distributed more fairly, and may give those who bear the greatest costs a say in that distribution.

received that the question of trust (O'Neill 2002) has resurfaced with such force in relation to major public data initiatives in health care (Caldicott Committee 2013).

6 Conclusion

I began with the difficulty of defining 'health data' and I end with the increasing ambiguity of health systems, in part because of the effects that adopting the technologies of data have themselves wrought on health systems or, at any rate, the developments that they have enabled (Beck 1992). Advances in information technology and data science are not peculiar to biological research and health care, although some fields, like bioinformatics, address these directly. The Nuffield report proposes an approach to the practical challenges of reconciling coincident interests in the use of data in biomedical research and health care, but one that is potentially of wider application to cases in which complex and interdependent interests are at stake. The *dynamic* and *cooperative* approach is appropriate to cases in which the definition of a data initiative may fluctuate and engage different and differing interests. It is especially relevant in non-zero-sum cases, such as health care and biomedical research, where the public interest is at stake, where optimising the design and performance of a system is itself a matter of public moral significance.

References

Bateson, Gregory. 1972. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. San Francisco: Chandler.

Beck, Ulrich. 1992. *Risk society: Towards a new modernity*. London: SAGE Publications Ltd.

Bennett, Craig M., Abigail A. Baird, Michael B. Miller, and George L. Wolford. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf. Accessed 11 June 2015.

Business, Innovation and Skills, UK Department for. 2011. *Strategy for UK life sciences*. London: Department for Business, Innovation & Skills.

Caldicott Committee. 2013. Information: to share or not to share? The information governance review. https://www.gov.uk/government/publications/the-information-governance-review. Accessed 5 Oct 2015.

Daniels, Norman, and James Sabin. 1997. Limits to health care: fair procedures, democratic deliberation and the legitimacy problem for insurers. *Philosophy and Public Affairs* 26(4): 303–350.

Faden, Ruth R., Nancy E. Kass, Steven N. Goodman, Peter Pronovost, Sean Tunis, and Tom L. Beauchamp. 2013. An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics. *Hastings Center Report* 43(s1): S16–S27.

Genomics England Ltd. 2015. Genomics England and the 100,000 Genomes Project. http://www.genomicsengland.co.uk/the-100000-genomes-project/ Accessed 24 June 2015.

Habermas, Jürgen. 1990. *Moral consciousness and communicative action*, trans. C. Lenhardt and S. Weber Nicholsen. Cambridge: Polity.

Kaye, Jane, Edgar A. Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. 2014. Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics* 23(2): 141–146.

Laurie, Graeme, Kerina H. Jones, Leslie Stevens, and Christine Dobbs. 2014. A review of evidence relating to harm resulting from uses of health and biomedical data. http://nuffieldbioethics.org/project/biological-health-data/evidence-gathering/. Accessed 24 June 2015.

Manson, Neil, and Onora O'Neill. 2007. *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work and think*. London: John Murray.

Nuffield Council on Bioethics. 2012. *Emerging biotechnologies: technology, choice and the public good*. London: Nuffield Council on Bioethics.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: Ethical issues*. London: Nuffield Council on Bioethics.

O'Neill, Onora. 2002. *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

Parker, Michael. 2007. Deliberative bioethics. In *Principles of health care ethics*, ed. Ashcroft Richard, Dawson Angus, Draper Heather, and John McMillan. Chichester: Wiley.

Prainsack, Barbara, and Alena Buyx. 2011. *Solidarity: Reflections on an emerging concept in bioethics*. London: Nuffield Council on Bioethics.

Prainsack, Barbara, and Alena Buyx. 2013. A solidarity-based approach to the governance of research biobanks. *Medical Law Review* 21: 71–91.

Taylor, Mark J. 2012. *Genetic data and the law: A critical perspective on privacy protection*. Cambridge: Cambridge University Press.

Taylor, Mark J., and Natasha Taylor. 2014. Health research access to personal confidential data in England and Wales: Assessing any gap in public attitude between preferable and acceptable models of consent. *Life Sciences, Society and Policy* 10: 15.

Vertesi, Janet. 2014. My experiment opting out of big data made me look like a criminal. *Time*. http://time.com/83200/privacy-internet-big-data-opt-out/. Accessed 1 July 2015.

Wanless, Derek. 2002. *Securing our future health: Taking a long-term view. Final report*. London: HM Treasury.

Wanless, Derek, John Appleby, Anthony Harrison, and Darsham Patel. 2007. *Our future health secured? A review of NHS funding and performance*. London: King's Fund.

Willetts, David. 2013. *Eight great technologies*. London: Policy Exchange.

The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts

Brent Daniel Mittelstadt and Luciano Floridi

Abstract The capacity to collect and analyse data is growing exponentially. Referred to as 'Big Data', this scientific, social and technological trend has helped create destabilising amounts of information, which can challenge accepted social and ethical norms. Big Data remains a fuzzy idea, emerging across social, scientific, and business contexts sometimes seemingly related only by the gigantic size of the datasets being considered. As is often the case with the cutting edge of scientific and technological progress, understanding of the ethical implications of Big Data lags behind. In order to bridge such a gap, this article systematically and comprehensively analyses academic literature concerning the ethical implications of Big Data, providing a watershed for future ethical investigations and regulations. Particular attention is paid to biomedical Big Data due to the inherent sensitivity of medical information. By means of a meta-analysis of the literature, a thematic narrative is provided to guide ethicists, data scientists, regulators and other stakeholders through what is already known or hypothesised about the ethical risks of this emerging and innovative phenomenon. Five key areas of concern are identified: (1) informed consent, (2) privacy (including anonymisation and data protection), (3) ownership, (4) epistemology and objectivity, and (5) 'Big Data Divides' created between those who have or lack the necessary resources to analyse increasingly large datasets. Critical gaps in the treatment of these themes are identified with suggestions for future research. Six additional areas of concern are then suggested which, although related have not yet attracted extensive debate in the existing literature. It is argued that they will require much closer scrutiny in the immediate future: (6) the dangers of ignoring group-level ethical harms; (7) the importance of epistemology in assessing the ethics of Big Data; (8) the changing nature of

This chapter is re-printed with the permission of Springer. The chapter was previously published in *Science and Engineering Ethics*: Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." *Science and Engineering Ethics*, 22(2): 303–341. doi:10.1007/s11948-015-9652-2. Page numbers from the original publication should be used when citing. Appendices excluded, please see original publication.

B.D. Mittelstadt (✉) • L. Floridi
Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford OX1 3JS, UK
e-mail: brent.mittelstadt@oii.ox.ac.uk; luciano.floridi@oii.ox.ac.uk

B.D. Mittelstadt, L. Floridi (eds.), *The Ethics of Biomedical Big Data*,
Law, Governance and Technology Series 29, DOI 10.1007/978-3-319-33525-4_19

fiduciary relationships that become increasingly data saturated; (9) the need to distinguish between 'academic' and 'commercial' Big Data practices in terms of potential harm to data subjects; (10) future problems with ownership of intellectual property generated from analysis of aggregated datasets; and (11) the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources. Considered together, these eleven themes provide a thorough critical framework to guide ethical assessment and governance of emerging Big Data practices.

1 Introduction

The amount of data being amassed by humanity is growing exponentially (Bail 2014, p. 465). Digital technologies, including online services and emerging ubiquitous computing devices, can track behaviour to a greater degree than ever possible (Markowetz et al. 2014). At the same time, policies as well as ethical, social and legal understanding of such "technological capabilities to merge, link, re-use and exchange data" lag behind the growth of technical capacities for data storage and analysis (Collingridge 1980; Safran et al. 2006, p. 6), the potential benefits of which are already being hailed in mass media (Markowetz et al. 2014, p. 407). Big Data (see below) has contributed to the definition of modern life as the 'information age'.

The technologies producing and processing data "provide destabilising amounts of knowledge and information which lack the regulating force of philosophy which . . . ensures that institutions remain rational." (Berry 2011, p. 8). New examples of the problems faced by Big Data systems appear regularly in mass media. Facebook's Beacon software, which was rolled out in 2007 to connect automatically external purchases to Facebook profiles, provided an early example of the ethically problematic nature of linking datasets. The service, intended to improve personalised advertising, inadvertently revealed sensitive characteristics of a person's life such as sexual preference (Oboler et al. 2012, p. 7). Big Data can similarly ground controversial forms of research, as demonstrated by the much discussed Facebook 'emotional contagion study' in 2012 (Schroeder 2014) A more recent and as-of-yet comparatively uncontroversial example are location awareness systems such as Apple's iBeacon software which connects information from a user's Apple profile to in-store systems and advertising boards, allowing for a 'personalised' shopping experience and tracking of (profiled) customers within physical stores (Apple 2014). Such systems effectively link online and offline personalities, supporting an "onlife" environment that may become invasive (Floridi 2014a).

As shown by these attention-grabbing examples, increasing interconnectivity in data-rich contexts can challenge accepted social and ethical norms. Practices centred on the mass curation and processing of personal data can quickly gain a negative connotation which, in a way similar to what has happened in the public debate over genetically modified organisms (cf. Devos et al. 2008), places potentially beneficial applications at risk through association with problematic applications. A 'whiplash

effect' can occur, by which overly restrictive measures (especially legislation and policies) are proposed in reaction to perceived harms, which overreact in order to re-establish the primacy of threatened values, such as privacy. Such a situation may be occurring at present as reflected in the debate on the proposed European Data Protection Regulation currently under consideration by the European Parliament (Wellcome Trust 2013), which may drastically restrict information-based medical research utilising aggregated datasets to uphold ethical ideals of data protection and informed consent (see Sect. 4.2.1).

Ethical foresight may reduce the probability of 'regulatory whiplash' by informing public debate through improved understanding of the 'moral potential' of emerging technological applications and data practices. To contribute to this process, a systematic and comprehensive review of academic literature discussing the ethical implications of Big Data was conducted to identify the issues of emerging importance for this novel form of data curation and analysis. Section 2 provides a brief background on 'Big Data' as a concept. Section 3 describes the methodology of a systematic and comprehensive review of academic literature discussing the ethics of Big Data, before presenting a narrative synthesis of the results in Sect 4. Shortcomings and further relevant issues not currently addressed sufficiently in the literature are then highlighted in Sect. 5, before concluding by reflecting on directions for further research in Sect. 6.

2 Background

'Big Data' covers a vast variety of phenomena focused on the analysis of large datasets. Data types and applications can be found in areas such as intelligence analytics (Mahajan et al. 2012), behaviour and preference modelling (Coll 2014, p. 1257; Lomborg and Bechmann 2014), sustainability studies (Mahajan et al. 2012), online and offline commerce, biomedical research and healthcare, and various other forms of scientific and social research and commercial pursuits (Costa 2014; Markowetz et al. 2014) based around mining vast datasets (Bail 2014, pp. 466–7). Data can be quantitative and textual, much of which is now user-generated via online behaviour that is revealing in terms of personal preferences and behaviours (Puschmann and Burgess 2014, p. 1694). The perceived value of such practices is variable, and may stem from characteristics such as the ability to collect data for research 'unintrusively' or perhaps, covertly (e.g. Lomborg and Bechmann 2014, p. 256), to track and profile fine-grained behaviours, preferences and other characteristics (e.g. sexual orientation or political opinions) of individuals (Coll 2014, p. 1257; Mahajan et al. 2012; Pariser 2011), to predict future behaviour (as used in law enforcement or credit, insurance and employment screening); or more broadly to search for connections across vast datasets for a variety of research purposes (Floridi 2012).

Against this broad context, biomedical Big Data has gained significant attention due to a combination of two factors. On the one hand, there is the huge potential to

advance the diagnosis, treatment, and prevention of diseases as well as foster healthy habits and practices (Costa 2014). On the other hand, there is the obvious, inherent sensitivity of health-related data and the implicit vulnerability and needs of those potentially requiring treatments (Pellegrino and Thomasma 1993). Academically and commercially[1] valuable biomedical big data can exist in many forms, including aggregated clinical trials (Costa 2014), genetic and microbiomic sequencing data[2] (Mathaiyan et al. 2013; McGuire et al. 2008; The NIH HMP Working Group et al. 2009), biological specimens, electronic health records and administrative hospital data.[3] Such data can be held in biobanks, cyberbanks and virtual research repositories[4] (Costa 2014, p. 436; Currie 2013; Majumder 2005, p. 32). Compared with traditional forms of storage, such repositories tend to assemble aggregated datasets explicitly for research purposes with "virtually unlimited opportunities for data linkage and data-mining" (Prainsack and Buyx 2013, p. 73) due to the sheer scale of the datasets (Steinsbekk et al. 2013, p. 151).

Data can also be generated explicitly or covertly via social media applications and health platforms[5] (Costa 2014; Lupton 2014, p. 858), emerging 'personal health monitoring' technologies (Mittelstadt et al. 2011, 2013) including wearable devices (Boye 2012), home sensors (Niemeijer et al. 2010) and smart phone applications, and online forums and search queries. The latter, for example, enable public health and outbreak tracking[6] (Butler 2013; Costa 2014, p. 435). Other data come from 'data brokers' which collect, process, store and sell intelligence based on a variety of medical and health-related data sourced from social media, online purchases, insurance claims, medical devices and clinical data provided by public health agencies and pharmacies, among others (Terry 2012, 2014).

Analysis of these data types can be undertaken for numerous purposes, including development of clinically useful predictive models (Choudhury et al. 2014, p. 3), longitudinal and cross-sectional effectiveness and interaction studies of

[1]For example, the identification of the presence of diabetes can support targeted marketing (Terry 2012, p. 392).

[2]For an overview of sample companies providing such services, see Costa (2014).

[3]In some contexts, such as the USA under HIPAA, administrative data will be afforded less protection than genomic and similar biobank data despite possessing similar capacities for revealing sensitive aspects of a person's health. This may be due partly to the possibility of removing identifiers from administrative data without 'ruining' the data (Currie 2013) as is an apparent limitation with anonymisation of genomic data (Hansson 2009, p. 10).

[4]These forms of biomedical data are incredibly varied and complex, consisting of data produced from a wide variety of sources, including "laboratory auto-analyzers, pharmacy systems, and clinical imaging systems . . . augmented by data from systems supporting health administrative functions such as patient demographics, insurance coverage, financial data, etc . . . clinical narrative information, captured electronically as structured data or transcribed 'free text' . . . electronic health records" to name but a few (Safran et al. 2006, p. 2).

[5]For instance, Facebook has recently announced plans for "support communities" and "preventative care applications" (Reuters 2014), while Google and Apple have recently released platforms for health and fitness data aggregation (Google Fit and Apple HealthKit/ResearchKit).

[6]However, the efficacy of such platforms remains questionable (Butler 2013).

pharmaceuticals (Tene and Polonetsky 2013, p. 246), and long-term 'personal health monitoring' (Boye 2012; Mittelstadt et al. 2014; Niemeijer et al. 2010). Broadly, these data may foster understanding of health disorders and the efficiency and effectiveness of treatments and health systems and organisations. They also create repositories for public health and information-based research (Safran et al. 2006, p. 2; Steinsbekk et al. 2013, p. 151). With that said, clinical applications are not guaranteed (Lewis et al. 2012). While promising on many fronts, biomedical Big Data, and the findings derived from it, may raise a host of ethical concerns stemming from the sensitivity of data being manipulated and the seemingly limitless potential uses and repurposing, and implications of data that concern individuals as well as groups.

3 Methodology

In order to understand what ethical issues have already been identified and discussed in the context of Big Data, a comprehensive and systematic meta-analysis of academic literature was conducted in October 2014. Six databases were searched (Web of Science, Scopus, Global Health, Philpapers, PubMed and Google Scholar) to identify literature discussing ethical aspects of Big Data. Search terms (with wildcards) were chosen to limit the review to articles explicitly mentioning 'Big Data' and ethics or morality, rather than searching for individual related concepts such as 'biobanks' or 'informed consent', thus allowing for a comprehensive review of Big Data literature. In recognition of the prevalence of biomedical applications, searches were divided between 'Big Data' and 'biomedical Big Data' to facilitate comparison. A breakdown of the search by database, search terms and results returned can be found in Table 1.

The title and abstract of each returned article was reviewed by the authors to determine relevance. Inclusion was based solely on the discussion of ethical issues in the article, with the goal of identifying themes in the literature. Limitations were not placed on the quality or length of the discussion, but rather on the mere presence of ethical concepts and issues. Further sources were also located through hand-searching and backtracking of citations provided within the reviewed articles.

The search was limited to English language articles. Although most of the reviewed literature consisted of peer-reviewed journal articles, other types of publications including commentaries, working reports, white papers and scientific books were also located. Date restrictions were not enforced.

3.1 Data Analysis

Each article was analysed and key passages highlighted for further interpretation and grouping into themes existing across multiple sources. These themes were allowed

Table 1 Search queries

Database	Search string	Returned
Ethics of biomedical big data		
Web of Science	**TOPIC**: ((ethic* OR moral*) (health* OR *medic* OR bio*) "big data")	23
Scopus	**TITLE-ABS-KEY**: ((ethic* OR moral*) AND (health* OR *medic* OR bio*) AND "big data")	18
Global Health	"Big data" AND (ethic* OR moral*)	145
PubMed	"Big data" AND (ethic* OR moral*)	19
Ethics of big data		
Web of Science	**TOPIC**: ((ethic* OR moral*) "big data" NOT (*medic* OR health* OR bio*))	18
Scopus	**TITLE-ABS-KEY**: ((ethic* OR moral*) AND "big data" AND NOT (*medic* OR health* OR bio*))	28
Philpapers	"Big data"	19
Google Scholar	"Big data" ethics OR ethical OR ethic OR moral OR morality OR morals	50*
Philosopher's Index	"Big data"	6

**11000 returned, first 50 reviewed*

to emerge from the literature rather than starting from a pre-defined theoretical framework. However, the review was intended to address two questions:

1. How is 'Big Data' conceptualised within discussions of its ethics?
2. What types of ethical issues are raised by Big Data?

To start, phrases and passages were highlighted that appeared to refer to ethical issues or concepts, understood as areas of 'right' and 'wrong' or the clash of competing values or normative interests among stakeholders. Highlighted segments were then coded to reflect the author's interpretation of the text (cf. Gadamer 2004; Patterson and Williams 2002). Similar codes were then grouped and assigned to ethical themes. Once themes had emerged from the literature, a second systematic analysis was performed using the NVivo 10 software package. All sources were re-checked via text search for the presence of the themes that emerged. The following terminological convention was adopted in discussing 'Big Data stakeholders' below: 'data subject' refers to the individual described by the data, 'data custodian' refers to any individual or organisation responsible for hosting or archiving the data in either its individual or aggregated form, and 'data analyst' refers to any individual or organisation analysing the data, but not necessarily hosting it.

4 Results

A total of 365 non-unique sources were identified for review across the databases, with 78 title/abstract combinations reviewed in full. Rejected sources were either off-topic or duplicates as determined by assessing the title and, in some cases,

abstract. Once fully reviewed, a further ten were excluded for being off-topic, leaving 68 sources that met the inclusion criteria of explicitly discussing ethical aspects of Big Data.

In terms of the types of Big Data discussed, 36 sources primarily addressed 'biomedical' data, or data with a health or medical connotation. The remaining 32 sources primarily discussed non-medical types of Big Data, referred to as 'general' data. For article types, 43 described original research or in-depth analyses of 'peer review' quality, while 25 were 'commentary' sources, which included consultancy documents, editorials and opinion pieces, section introductions and other short pieces that do not always require peer review, empirical research, or extensive referencing. Only 7 of the 68 sources included empirical research. Finally, 23 of the 68 sources featured an in-depth discussion of ethics or ethical aspects of Big Data.

The results of the meta-analysis are presented as a narrative overview, which highlights and comments upon key themes and topics in the literature. This overview is intended to address the two aforementioned questions in order to provide a starting point and reference for future discussions concerning the development, regulation, and ethical evaluation of Big Data practices.

4.1 Conceptualising Big Data

To identify how Big Data is conceptualised in ethics literature, it is necessary to consider how it is defined and which applications are discussed as posing potential ethical harms. A commonly accepted definition of 'Big Data' was not reflected in the literature. While definitions vary, some common characteristics and frameworks can be found. The first and most influential definition of Big Data was provided by Laney (2001) in terms of three dimensions: (1) Volume, or the scale of data; (2) Velocity, or the analysis of streaming data; and (3) Variety, or different forms of data. Later, another V was added: (4) Veracity, or the uncertainty of data (IBM 2014), giving rise to an influential framework (Andrejevic 2014; McNeely and Hahm 2014; Nunan and Di Domenico 2013). Accordingly, Big Data is unique in terms of the size and "speed of data generation and processing and the heterogeneity of data that can be dumped into combined databases" (Andrejevic 2014, p. 1676). Aggregation is justified by the idea that "things can be learned from a large body of data that cannot be comprehended from smaller amounts," revealing the implicit link between Big Data and complexity (McNeely and Hahm 2014, p. 305).

Broadly, Big Data can refer to (1) the *process* of analysing 'big' data sets, and (2) the *datasets* themselves. 'Big' can be defined variably in terms of quantities of electronic size (gigabytes, terabytes, petabytes, etc.), entries, individuals or events represented by the data, or alternatively in relation to the techniques and technologies currently available for analysis. The latter approach defines 'big' in procedural rather than quantitative terms, by connecting the size of the dataset to its complexity, understood in terms of the computational or human effort necessary

for analysis (e.g. Costa 2014; Dereli et al. 2014; Fan and Bifet 2013; McNeely and Hahm 2014; National Science Foundation 2014; Terry 2012, p. 389). In other words, the data is 'Big' because it is difficult to sort and analyse with existing computing technologies.

While helpful for bridging the space between analysis processes and datasets, this approach suggests data that is 'Big' now may not be so in a year or a decade due to advances in computing technology and analysis procedures (Floridi 2012; Liyanage et al. 2014, p. 27). Although not semantically problematic (as adjectives describing technology tend to be relative, e.g. fast internet 10 years ago is slow internet today), this nevertheless poses a technological solution to an epistemological query by making the definition of 'Big Data' relative in relation to technical and analytical capacities. 'Big Data' becomes data that is difficult to analyse due to its size and complexity. This also suggests that more or better computing will enable us to 'get ahead' of the data and analyse all of it meaningfully again, as we did prior to the current era of Big Data. However, the exponential growth of data (Bail 2014, p. 465) suggests this is unlikely to occur, a point that further reinforces the view that Big Data describes a break with prior practice. Explicit consideration of historical context reduces the fluidity of the definition; in other words, labelling a study as 'Big Data' recognises the technical and analytical barriers faced at the time it occurred. Such fixed labelling may be important in ex-post ethical analysis (see Sect. 5.5).

Recognising these implications of a purely technical definition, it may be helpful to consider also the perceived value of Big Data as suggested in the types of analysis it allows. Boyd and Crawford (2012, p. 663) suggest Big Data is valuable due to the "capacity to search, aggregate, and cross-reference large data sets." Similarly, according to Floridi (2012), a unique feature of Big Data is the possibility of identifying small patterns and connections in quantitatively large (and often aggregated) datasets. 'Small patterns' refer to connections between entries within the dataset, meaning connections are found within a subset of entries in a much larger dataset.

4.1.1 State of Deployment

As an emerging concept, defining fixed boundaries or practices which are 'Big Data' is perhaps impossible. Despite this, to avoid unbridled speculation over future ethical implications of Big Data practices, it is useful to consider first the state of development and deployment of different practices as reflected in the literature. Figure 1 shows a timeline describing the current state of various Big Data applications and practices categorised according to likelihood for deployment or commercial/research applications, ranging from: (1) real world, or currently in use; (2) on the horizon, with high likelihood of materialising due to the existence of the necessary technologies or data and empirically demonstrable motivation for use; or (3) not currently possible, meaning deployment may theoretically be possible but of limited likelihood or frequency due to access limitations or technical, ethical and other constraints. The boxes were populated to reflect Big Data technologies and

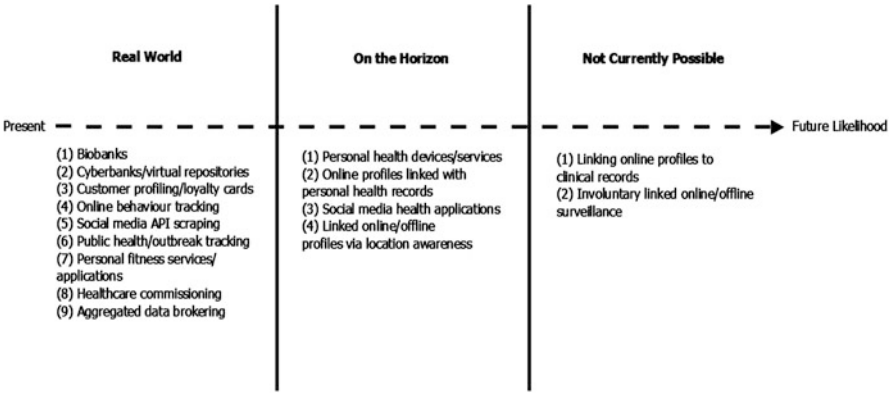

Fig. 1 Estimated timeline of big data applications

practices currently in use, on the horizon in terms of research and development, and imagined but not yet possible. Each category and its position is reflected in the reviewed literature: for example, biobanks, cyberbanks, loyalty cards and personal health monitors are all widespread technologies, enabling large-scale data collection and assessment. At the other end of the scale, linking of offline medical records with online profiles (such as a Facebook user account) is currently seen as potentially valuable but unlikely due to ethical concerns, which would likely also preclude linked online/offline tracking of individuals without consent. With that said, the table is intended as an informed estimate of Big Data deployment rather than a description of the state-of-the-art. The latter was not deemed possible given the varied age and empirical evidence-base of the reviewed papers. Furthermore, technical development likely precedes commentary on social and ethical aspects (cf. Collingridge 1980; Moor 1985). Importantly, the figure reflects the likelihood of *widespread* deployment, meaning devices and uses currently in development, while testing or limited public release are considered 'on the horizon' rather than 'real world', e.g. profiles linked via location awareness, see (Apple 2014).

4.2 Ethical Themes

Through content meta-analysis five major ethical themes emerged from the literature. Interpretation and designation of themes were discussed and agreed upon by the authors. Although the ethical themes emerged according to frequency, the overview does not merely highlight this frequency. Rather, the results discussed in the following sections were chosen for one of four reasons: (1) to draw attention to common interpretations of ethical themes and concepts, (2) to emphasise individual cases and issues that reveal unique ethical aspects of Big Data, (3) to highlight studies with an in-depth analysis of ethical concepts and issues, and (4)

Table 2 Ethical themes

Theme	N. of sources
Informed consent	34
Privacy	44
Anonymisation	20
Data protection	14
Ownership	12
Epistemology	14
Power/control (*big data divide 1*)	22
Digital divide (*big data divide 2*)	22

to identify gaps in the discussion in need of further research. The presentation of results therefore focuses on the authors' analysis and interpretation of the literature (Table 2).

4.2.1 Informed Consent

Half of the literature addresses issues of informed consent. The concept (cf. Angrist 2009; General Medical Council 2008) does not cleanly transfer to research involving Big Data for a variety of reasons. Historically, consent is taken for participation in a single study, not covering unrelated investigations resulting from sharing, aggregating, or even repurposing data within the wider research community (Choudhury et al. 2014, p. 4). This form of consent is problematic because Big Data is intended by design to reveal unforeseen connections between data points. This means that both what the data reveals about the subject and its utility in future research present greater uncertainty than normal at the time of consent. For example, secondary effects of pharmaceuticals can be identified by comparing data not only from multiple clinical trials, but 'informal sources' as well, such as incidental self-reporting via social media and search engine queries. In this type of research the connections that can be revealed through linking multiple data sets cannot be accurately predicted prior to carrying out the research. As a result, 'consent' cannot be 'informed' in the sense that data subjects cannot be told about future uses and consequences of their data, which are unknowable at the time the data is collected or aggregated.

Recognising this, calls to reform single-instance consent based on the belief that it is a barrier to 'necessary' research and innovation can be found in the debate (Larson 2013). 'Broad' and 'blanket' consent mechanisms, which pre-authorise future secondary analyses, are sometimes used in place of single-instance consent (Clayton 2005; Ioannidis 2013), if only due to the impracticality of renewing consent for each new analysis (Currie 2013; Lomborg and Bechmann 2014, p. 262). Such barriers often lead to biobanks employing a broad type of consent covering all future research activities. However, this approach has been recognised to limit the autonomy of data subjects (Master et al. 2014, p. 1). Tiered consent can also

be used, which provides the data subjects with options for permitting specific uses of their data – for example to allow the data to be used in cancer research but not in genomic research – or to require specific re-consent for future uses rather than blanket consent for all potential uses (Majumder 2005, p. 33). Exclusion clauses can be used for a 'line-veto' type of tiered consent, which can increase confidence in data subjects that custodians are actually respecting their beliefs and values as translated into prohibitions of specific re-uses (Master et al. 2014). Where such formats are used, governance mechanisms, such as review councils and committees, help distinguish '*bona fide*' and problematic requests for access to data.[7] Note that differences in broad or blanket consent across national borders may also complicate sharing and re-use of data, an issue that is likely to become more pressing in the future, especially at the European level.

The impossibility of certainty concerning future uses of data highlights a key aspect of 'Big Data', namely the desire for openness and creativity in identifying novel connections between data sets. For data collected explicitly for aggregation into a 'Big Data set', the openness of the format does not create difficulties, although open data sharing may require a global type of consent, due to data travelling across the political and electronic borders of institutions and nations (Majumder 2005, p. 33). The same cannot be said for historical data for which consent was granted for a specific purpose. By the ideal of explicit single-instance consent such data should not be used without explicit consent for secondary uses by the person providing it. However, obtaining such consent years after a trial has been conducted can be very difficult, if not impossible (Clayton 2005; Wellcome Trust 2013). A tension therefore exists between the potential benefits of 'Big Data' analysis of historical datasets and the need for consent. Technical compromises are not obviously the solution – for instance, anonymising the data may not be sufficient to eliminate the need for consent due to the possibility of re-identification (Mello et al. 2013, p. 1653; Tene and Polonetsky 2013, p. 251; Terry 2014, p. 837). The temptation to conduct research on datasets beyond the boundaries of the initial consent agreement is at the heart of Big Data—in some cases, not only is consent no longer sought, but instead a 'duty to participate' in secondary research, thought of as 'research that is not research', is described (Ioannidis 2013, p. 40).

These two issues of consent are clearly applicable to data obtained from scientific (in particular, biomedical) research for which informed consent has long been standard. However, new issues are raised by the collection and analysis of data from potentially 'unwilling' participants, for example data scraped from social media platforms, smart phone applications, or open web forums (e.g. Krotoski 2012; Lomborg and Bechmann 2014; Markowetz et al. 2014). Terms of Service and other end-user agreements governing the usage of these applications tend to allow for collection, aggregation and analysis of such data. Indeed, social networking platforms such as Facebook and Twitter rely heavily upon advertising revenues

[7] See for example the UK Biobank Ethics and Governance Council: http://www.egcukbiobank.org.uk/.

generated through just such practices. However, as social scientific and other forms of research begin to utilise data collected from the unwilling or uninformed (cf. Enjolras 2014; Lazer et al. 2009), the lack of an explicit *informed* consent mechanism in end-user agreements gives cause for concern (Fairfield and Shtein 2014), even when 'participants' are 'de-identified' (Ioannidis 2013).

Modifying Consent for Big Data

An unintended consequence of overly restrictive data protection and distribution policies is that a barrier can be erected to sharing data between researchers that is otherwise acceptable to data subjects (Choudhury et al. 2014, p. 5). In doing so, researchers may be missing opportunities to derive valuable information and innovations from the samples and data offered by research participants. Such a situation is currently materialising in relation to the European Data Protection Regulation under debate in the European Parliament. The regulation may severely restrict 'Big Data sciences' by requiring "specific, informed and explicit consent" for each instance of processing or analysis of 'personal data' (as specified in Article 83). The potential existence of Big Data research is therefore currently in jeopardy in Europe due to consent requirements (EURORDIS 2013; Wellcome Trust 2013). As argued by a consortium of biobanks, research councils and trusts (Wellcome Trust 2013), such a requirement creates barriers for 'big' datasets containing data from (hundreds of) thousands of individuals that are impractical and perhaps impossible to overcome in practice; indeed, most such repositories use blanket, broad or tiered consent.

If the Regulation were passed in its current form, each request by researchers for access to records held by biobanks would require re-contacting and re-consenting the data subject (EURORDIS 2013). In many cases, this will be impossible due to changes of contact details or death. Even where possible, it presents a significant financial and bureaucratic barrier to research (Wellcome Trust 2013). In this context, explicit, single-instance informed consent is causing rather than solving ethical problems by creating barriers for legitimate forms of research in addition to those rightly viewed as challenging, thus preventing researchers from advancing scientific knowledge, from deriving beneficial applications, and more generally from fulfilling the moral obligation to data subjects that have volunteered their time, bodies, and data for research.[8]

[8]By some accounts moral obligations exist for medical research. As suggested by accounts of solidarity-based governance of biomedical Big Data (e.g. Prainsack and Buyx 2013), patients may have a moral duty to participate in research due to the value generated through advances in medical knowledge and treatments (Harris 2005; Schaefer et al. 2009). As participation inherently includes risks, researchers may similarly have a moral obligation to minimise risks as far as possible by extracting maximum value from existing datasets through re-purposing and aggregation (Currie 2013; Harris 2005).

For the problems faced in Europe, these difficulties may be solved by introducing a number of clarifications and modifications to Article 83 of the Regulation, focusing mainly on exemption of pseudonymized data from the need for 'explicit' consent without which a "disproportionate regulatory burden" would be created, equivalent to that governing identifiable data (EURORDIS 2013; Wellcome Trust 2013, p. 2). More broadly, numerous approaches to consent have been proposed to overcome such barriers in purely information-based research, which re-use existing datasets (e.g. Prainsack and Buyx 2013; Rothstein and Shoben 2013; Schadt 2012). These 'fixes' tend to eliminate the need for consent to some degree in two main ways: pragmatically, by relying upon altruism, or substantively, by using an 'opt-out' approach, emphasising solidarity (see for example Prainsack and Buyx 2013) or the public good rather than individual autonomy (Rothstein and Shoben 2013).

Concerning pragmatic solutions, in the context of genomic sequencing, research is sometimes restricted to "information altruists" (Choudhury et al. 2014), or individuals willing to openly share their data (and sometimes, identity) on the basis that they possess the social status or economic resources to be sufficiently protected from future discrimination or harmful consequences. 'Radical honesty' models are similar, through which individuals volunteer de-identified genetic information for public sharing (Hayden 2012). Another approach is to establish "honest broker" and "stewardship" consent models by which impartial third parties mediate broad consent agreements to protect the interests of data subjects (Choudhury et al. 2014, p. 7; Goodman 2014). Emphasising professionalism or enacting punitive measures for misuse of data can shift some of the burden to researchers benefiting from access to the data and promote feelings of responsibility to data subjects (Fairfield and Shtein 2014). The hope here is that forbidding unacceptable forms or research, such as re-identification of anonymised data, will minimise potential negative impacts on data subjects (Hayden 2012, p. 314).

Concerning substantive solutions, an 'opt-out' approach to consent (e.g. Hoffman 2014; Rothstein and Shoben 2013; Tene and Polonetsky 2013; Terry 2012) should not be seen as ethically equivalent to informed consent. Opt-out consent models may take advantage of people in vulnerable moments (Hayden 2012), for example if consent is taken during a clinical encounter in which the data subject is seeking treatment (cf. MacIntyre 2007; Pellegrino and Thomasma 1993). However, the weaknesses of such approaches do not suggest that explicit consent for each instance of data use is the correct path either; rather, a revision of ethical standards which strikes a balance between the requirement for consent and the practical requirements of 'Big Data science' may be appropriate. Tene and Polonetsky (2013, p. 262) suggest as much in calling for debate on the "merits of a given data use" as a broader societal issue, wherein distinctions can be drawn between 'types' of data uses requiring full informed, opt-in, opt-out or no consent at all.

It may be possible to reduce or eliminate the need for consent by focusing on the concept of solidarity and the alleged reduced risks to data subjects in data-based research. Prainsack and Buyx (2013) suggest that a solidarity-based approach to biobank governance, focused on harm mitigation, can be used in place

of informed consent,[9] which recognises an empirically supported sentiment among the general public (in Europe) to want to participate in biobanking research (Kaye et al. 2012; Steinsbekk et al. 2013). Rather than tweaking consent as such, their approach seeks to re-define the relationship between biobanks and data subjects by emphasising the willingness to share data or assist others to support research and innovation (Prainsack and Buyx 2013, p. 74). In contrast to autonomy-based consent approaches, biobanks would instead model consent on solidarity by providing data subjects with a 'mission statement', information on potential areas of research, future uses, risks and benefits, feedback procedures and the potential commercial value of the data, so as to establish a "contractual" rather than consent basis for the research relationship (Prainsack and Buyx 2013, p. 84). Such an approach is claimed to be acceptable given the relatively low risks in genomic research. According to the authors, few examples of discrimination based on biobank-facilitated research exist and are incomparable in quality to the bodily harm possible in other types of medical research.[10] One of the most commonly cited fears of insurance discrimination based on disease susceptibility is dismissed due to the "limited predictive value" of genetic markers at the individual level (Prainsack and Buyx 2013, p. 79).

As a counterpoint to solidarity, the alleged 'responsibility' to give up consent rights to prevent "hindering progress" in scientific research and thus social good can be seen as an unethical burden placed on the individual (Crawford et al. 2014, p. 1666). Given the extensive uncertainty over what collected data may reveal in the future, eliminating or reducing the need for informed consent based on the solidarity of the general public cannot be accepted uncritically and seemingly without public debate, particularly if democratic ideals are valued. With that said, the acceptability changes drastically with timescale—possible implications decades in the future are unlikely to outweigh the potential benefits of data sharing now. Another possibility not relying on such a problematic form of responsibility may be to emphasise trust between governance bodies wherein data is shared only between 'trusted' bodies (cf. Hansson 2009, p. 9); however, this may only be a tenable alternative for research data repositories where the extent and initial purpose of data collected is known to data subjects.

[9]The shift to solidarity is also said to free up the "significant resources" currently spent on (re-)consenting procedures for primary and secondary uses of data held in biobanks for research, innovation and infrastructural improvements including interoperability between repositories (Prainsack and Buyx 2013, p. 80). This position rests on the assumption that significant resources are currently being spent on re-consent procedures in particular, which are a central concern for consent and Big Data (e.g. Wellcome Trust 2013), and that these resources would instead be spent on valuable research and structural improvements.

[10]The relative lack of reporting on harms stemming from abuses of biomedical data has been noted in a recent Nuffield Council report on the ethics of linking biomedical datasets for research (Nuffield Council on Bioethics 2015). The lack has been largely attributed to a lack of robust reporting mechanisms and empirical research on underreporting, with most cases coming from anecdotal accounts and notable media stories. As a result a lack of evidence of harms should not be considered evidence for a lack of harms.

4.2.2 Privacy

Unsurprisingly, privacy features very frequently in the literature, often in parallel with anonymisation and confidentiality. Commentary pieces often address privacy issues of Big Data (e.g. Craig 2011; Goodman 2014; Schadt 2012), presumably due to the prevalence of the concept in international legislation and related discussions in applied ethics. In the reviewed literature, numerous concerns were described in terms of privacy, some of which relate to alternative concepts such as autonomy or freedom of information. Links are frequently made with confidentiality, understood as "the duties that accompany the disclosure of non-public information within a fiduciary, professional or contractual relationship" (Majumder 2005, p. 33). Others discussed privacy in terms of the 'invasiveness' of Big Data analysis. Invasiveness was connected in particular to analysis of combined data sets, particularly from geolocation and internet-based sources, even when such data is anonymised (e.g. Markowetz et al. 2014; Moore et al. 2013; Shilton 2012).

Where explicit theories and frameworks of privacy were applied, the OECD's Fair Information Principles and Nissenbaum's 'contextual integrity' (2004) were influential (see for example Andrejevic 2014; Helbing and Balietti 2011; Tene and Polonetsky 2013). Nissenbaum's context-sensitive approach to privacy norms is clearly relevant for emerging forms of participatory data generation such as social media, where data subjects may not be aware of the extent to which data can be publicly 'scraped' and analysed outside of the "highly context-sensitive spaces" in which it is created. Such uses may violate subjects' expectations of data privacy (Boyd and Crawford 2012, p. 673) and expose the data to acontextual interpretation (e.g. Andrejevic 2014, p. 1685). A conceptual link can be drawn to the distinction between 'being in public', in the sense that data communicated via the internet is publicly visible by default, and 'being public', or asserting one's agency purposefully to make something publicly known. This distinction is often ignored in Big Data (Boyd and Crawford 2012, p. 673), insofar as being able to do something becomes synonymous to being justified in doing it. To facilitate formation of realistic privacy norms in Big Data contexts it may be necessary to reinforce the distinction in digital spaces between 'being in public' and 'being public'. 'Offline' privacy barriers such as physical walls can be replaced by raising awareness among data subjects of the uncertain but broad value and seemingly limitless lifespan of the data outside of the original context in which it was authored. Awareness may inhibit authorship or dissemination of sensitive or particularly context-sensitive data.

Following from this, the scope of data being collected can also be conceived of as a privacy issue. Traditionally, data collection has been limited by human perception and cognition. However, with automated and autonomous collection by information technologies, the scope of data, as can be seen over the past two decades, has grown exponentially. More personal and highly detailed data can be collected and analysed than at any other time in history (Nunan and Di Domenico 2013, p. 5). This is a unique characteristic of the 'age of Big Data' (Andrejevic 2014; Puschmann and Burgess 2014). Furthermore, these data are designed to be stored in perpetuity, meaning that traditional limitations of memory no longer apply; data collected

today may, in theory, be equally accessible and of the same quality in the future. Other issues are now emerging as important for the preservation of data, such as the obsolescence of software, the presence of malware, and the potential fragility of physical supports. While not a privacy issue *per se*, extending the lifespan of data describing phenomena that would otherwise be forgotten does increase the risks that privacy violations may occur.

Anonymisation

Anonymisation and privacy were closely linked in the literature, wherein privacy concerns raised by Big Data practices can be addressed merely by removing identifying information. Anonymisation was frequently seen as the minimum requirement necessary to protect data subjects' privacy in aggregating data, despite the possibility of re-identification through cross-referencing with data concerning ethnic background, locational data, other metadata, health records or even small pieces of identified genetic data (Choudhury et al. 2014, p. 6; Hayden 2012, p. 313; Joly et al. 2012).

For medical research, data is often anonymised or de-identified to gain consent from data subjects and in accordance with data protection legislation. Beyond explicit 'biobanking' research, study data collected in this way populate Big Datasets. Interestingly, some authors hold that "when analysing data at an aggregate level, a waiver of consent" is acceptable (Krotoski 2012, p. 30), directly linking the acceptability of analysis and re-use of data to the issue of anonymisation (see section "Modifying Consent for Big Data"). However, this position is problematic because it portrays consent as a concept relevant only to the identifiable individual, whereas group-level harms from analysis of aggregated data are clearly possible (Floridi 2014b). Whereas traditionally research ethics has focused on the harms to individual participants, Big Data operates on and impacts groups (of anonymised individuals) (Fairfield and Shtein 2014). Where anonymised data subjects are grouped according to geographical, socioeconomic, ethnic or other characteristics, the anonymisation of individuals matters little if outcomes affect the groups to which they belong (Choudhury et al. 2014, p. 6). Problematic discrimination and stigmatisation of affected groups (see Sect. 5.1) is therefore a real risk (Docherty 2014), even in anonymised datasets. Such effects impact on all members of the community, not only those who gave consent (Fairfield and Shtein 2014, p. 45).

As shown at the group-level, anonymisation can be criticised when presented as a 'silver bullet' that avoids, or at least minimises, the risk of being 'singled out' for discrimination or preferential treatment (McGuire et al. 2012). An 'ethics of care' approach may be appropriate when working with Big Data collected from groups based on, for example, indigenous, demographic, ethnic or cultural features, to avoid possibilities of discrimination (Lewis et al. 2012, p. 3).[11]

[11]The applicability of theories on the ethics of care (e.g. Gilligan 1982; Noddings 2013; Slote 2007) to Big Data likely extend beyond discrimination against marginalised groups. For example, emphasising responsiveness and relationships between data subjects, custodians and analysts may

Data Protection

Current data protection legislation in the USA and EU may not protect all medically-relevant or health-related Big Data, or afford such data the protections granted to sensitive health data. As a result, usage of these data will largely be governed by the ethical systems and values governing particular databases or custodians (Liyanage et al. 2014, p. 33), such as institutional or ethical review boards. This situation is particularly concerning for privatized and internet-based health data sources, such as patient-driven databases (e.g. PatientsLikeMe) (Liyanage et al. 2014, p. 33), which are likely to be subjected to less stringent requirements when compared to biobanks and repositories of clinical trial data, where restrictions can be enforced by governance bodies.

4.2.3 Ownership

Ownership is a complex concept, as it can refer to rights regarding the redistribution and modification of data, along with benefiting from intellectual property and innovations developed from its analysis. Redistribution and modification of data may be restricted by the data 'owner' to maintain data integrity, while access is still allowed to data 'analysts' for analysis, innovation, and development of intellectual property. Different databases will have different restrictions in place. A key distinction is that there are two forms of ownership, as rights to 'control' data, and as rights to 'benefit from' data. The former of these two conceptions was primarily discussed in the literature, perhaps due to the nascent development of intellectual property and products from Big Data thus far.

Understood in terms of 'control', ownership grounds empowerment of data subjects through mechanisms to track and check the existence and manipulation of their data, which can help prevent "the existence of "secret" databases and leverage societal pressure to constrain any unacceptable uses" (Tene and Polonetsky 2013, p. 242). Here, a link can be seen between discrimination and surveillance (see Sect. 4.2.5). When the possibilities of re-identification and 'hidden' analysis exist, data subjects' control over uses of their data acquires greater importance (Choudhury et al. 2014, p. 6), as control allows subjects to restrict undesired uses. For biobanks, control is relevant to considering the permissibility of using research data for commercial pursuits as is made possible by allowing private and third party companies access (e.g. NHS England 2014). Permissibility can be described as an ethical issue when 'human dignity' precludes commodification of humans, or selling one's body or data describing one's body (Steinsbekk et al. 2013).

provide avenues for development of new privacy protection mechanisms and group-level ethics which acknowledge the network ethical effects possible through Big Data (see Section 5.1). While a full account of this and related topics concerning ethics of care goes beyond the scope of this paper, existing work on the applicability of the ethics of care to public health (e.g. Kass 2001) may provide a starting point for future enquiries.

Understood as a 'benefit', ownership can also require data custodians to enable data subjects to benefit and utilise Big Data for personal uses by being offered "meaningful rights to access their data in a usable, machine-readable format" (Tene and Polonetsky 2013, p. 242). Such steps allow subjects to find individual benefits from the data they produce and communities (or aggregated datasets) in which it resides (Lupton 2014, p. 866).

Accessibility in both contexts is not without risks and necessary limitations. For instance, providing data subjects with unrestricted access to raw data may be harmful in the sense that it is practically useless or open to misinterpretation without the presence of a trained clinician or analyst to explain its significance (Watson et al. 2010). Furthermore, revision rights open datasets to mistakes and inaccurate modification by data subjects, while not addressing questions of accuracy of interpretations or the completeness of the data representations.

4.2.4 Epistemology

Interestingly, despite the search being restricted to literature discussing 'ethics', a number of sources revealed a connection between the ethics and epistemology of Big Data. The connections stem from the perceived complexity of Big Data and the algorithms used to analyse it (Callebaut 2012, p. 70), which may exceed human comprehension – "the intelligent citizen cannot read the programs that run our data sets." In other words, "the natural world and its human observers are being ever more instrumented with intelligent machines ... as people we are, in Olga Kuchinskaya's memorable phrase, becoming our own data" (Bowker 2013, p. 170). The problem is not new: patients are usually unable to interpret radiographies, for example. But it has become more significant because of its size, opacity, and pervasiveness. Complexity now refers both to the inherent difficulty of analysing vast datasets, and to the complicated reasoning or rationale of the algorithms (or analytical processes) that make discoveries in Big Data. As a result, questioning the validity of relationships and findings based upon analysis of Big Data becomes increasingly difficult not just for the general public but also for experts, whose critical investigations may become comparable to questioning the outputs of a 'black box'.

Objectivity

The aforementioned complexity leads to several related problems. One is a tendency, particularly in mass media and industry, to view Big Data as 'objective'(Crawford 2013; Crawford et al. 2014) or as revealing objective truths without the need for human interpretation. This 'mythological' view of Big Data as the 'end of theory' creates ethical concerns regarding justification of increasingly pervasive and unbounded secondary manipulation and aggregation of data when Big Data practices are seen as the future of science and scientific discoveries. Data are (mistakenly) said to "speak for themselves", creating the possibility of science being

driven entirely by induction and reduction, or 'data-driven science' without a need for theory or hypotheses (Callebaut 2012, p. 70; Crawford et al. 2014; Fairfield and Shtein 2014, p. 47). For proponents, the "sheer abundance of information" is seen as providing a "degree of scientific authority" to Big Data practices, which can seemingly be used to explain any "natural and social phenomena" (Puschmann and Burgess 2014, p. 1691) because its meaning is "already there, just waiting to be uncovered" (Puschmann and Burgess 2014, p. 1699).[12]

This idea of data-led objective discoveries entirely discounts the role of interpretive frameworks in making sense of data which, according to non-objectivist ontologies (e.g. Gadamer 1976; Habermas 1984, 1985; Heidegger 1967; Schwandt 2000), is a necessary and inevitable part of interacting with the world, people and phenomena (and thus, data). In the increasingly abstract and complex practices that make up Big Data (Puschmann and Burgess 2014, p. 1697) "data is extracted, collected, cleaned and transformed, stored and managed, analysed, indexed and searched, as well as visualized" (Markowetz et al. 2014, p. 407). In unstructured searches for 'patterns in the data', 'noise' is eliminated as the dataset's boundaries are modified to facilitate the search (Puschmann and Burgess 2014, p. 1699). At each step the data undergoes a transformation by passing through an interpretive framework, yet custodians act as though it remains an objective analogue of reality. What is or may be relevant depends on the questions being asked, which in turn depend on the purposes for which the investigation is being developed. Only a clear understanding of the purposes can ground a rational determination of the levels of abstraction at which the data are queried. The need for human intelligence is actually increasing the more data become available, in order to know which sensible questions to ask and what answers actually make sense (Floridi 2008, 2013).

The tendency to rely on mere Big Data furthermore ignores the variable quality of datasets. For instance, electronic health records typically consist of data written by clinicians for clinical work without the interests of researchers, standardisation and interoperability in mind, while aggregation of observational data for purposes of identifying causal links is prone to selection, confounding and measurement biases (Hoffman and Podgurski 2013; Ioannidis 2013, p. 40). If data come to be processed automatically without "human checks" (Hoffman and Podgurski 2013, p. 56) or by algorithms beyond the capabilities of human understanding, the variable quality of the data undermines justification of the actions taken on their behalf.

Some may argue in favour for a distinction between 'objective' or 'raw' data about the physical world and necessarily subjective or interpretive data about human behaviour or social reality, which is nevertheless treated as similarly objective when labelled as 'Big Data' (cf. Lupton 2014, p. 859; Schroeder and Cowls 2014). However, regardless of the position taken on this issue, the key message is that the objectivity of Big Data describing social reality (which includes biomedical data) is

[12]With these tendencies noted, the capacity of Big Data to provide scientific explanations of particular types of social phenomena or human behaviours should not be rejected (e.g. Schroeder 2014).

often falsely represented by those treating data as inherently neutral and capable of explaining complex phenomena (e.g. 'data-driven science') without need of further contextual knowledge, meaning, or interpretation.

Context

When knowledge is seen to 'emerge' from 'raw' data, the need for understanding the contextual meaning or 'situatedness' of the data is seemingly dismissed. Even where the need is acknowledged, contextual understanding may be impossible in aggregated datasets; for instance, as recognised in cultural sociology, "big data often does not include information about the social context in which texts are produced" (Bail 2014, p. 477). Context and meaning may also be purposefully stripped from behaviour and actions, for instance when tracking behaviour via Big Data is viewed as a 'less biased' way to collect behavioural data compared to self-reporting and questionnaires (e.g. Markowetz et al. 2014, p. 406). For research studies, data is collected and interpreted in a particular way to "solve a specific problem, characterized by a limited focus and functionality," limiting possibilities for interoperability between 'stovepipes' or datasets (McNeely and Hahm 2014, p. 306). A tangible loss of methodological and scholarly context occurs through the aggregation of data unless extensive precautions are taken to preserve the 'assumptions' that helped generate the data: "the categories to be used in collecting data, the procedures for handling missing data, the specific subjects of data collection, the nature of the sampling methods used, and the means by which to construct and aggregate the data" (Busch 2014, p. 1730). In other words, aggregation obscures the complex methodological decisions and ontological assumptions that ground the research that produced the data to be aggregated. Busch (2014) describes the following characteristics of aggregated datasets which contribute to a loss of context:

- **Lossiness** – Aggregation, case construction, standardisation and simplification of data to enable cross-sectional analysis may 'lose' certain aspects of the phenomena studied.
- **Drift** – Phenomena change over time, but the data representing them does not. The same can be said for the methods underlying the primary data collection and analysis.
- **Distancing** – Large datasets facilitate identification of patterns or 'clarity' by distancing oneself from the phenomenon.
- **Layering** – A 'realist' ontology is pre-supposed in that an assumption is made that the relations underlying the phenomena will remain over time as the data is aggregated and manipulated. The 'situatedness' or contextual meaning of each phenomenon must be removed for the (data representations of the) phenomena to be treated as sufficiently similar for aggregation. Context is lost by reducing the phenomena to a set of variables: "those aspects of things that are not amenable to numerical or statistical analysis – that situate particular phenomena – are systematically downgraded or removed from consideration" (Busch 2014, p. 1735).

- **Errors** – How are errors within the dataset identified and addressed?
- **Standards** – "The process of creating uniformity through standardization," or fitting data to discipline conventions or categories, "may obfuscate phenomena of considerable importance" (Busch 2014, p. 1736).
- **Disproportionality** – Outlying data may be deleted or treated as 'errors' to enable simplification and standardisation of the dataset.
- **Amplification/Reduction** – Aspects of phenomena amenable to quantified measurement are amplified in importance, while those that are not are reduced.
- **Narratives** – Large datasets can hide the role of interpretation in seeing data *as* something and obscure alternatives to the preferred interpretation.

Contextual aspects which do not fit the structure or classification framework of a database appear to be (sometime irreversibly) lost in Big Data, in some cases through computer-led 'interpretation' of the data (Bowker 2013, p. 170), particularly in social research. This can be referred to as a "signal problem" wherein data is treated as an accurate representation of social reality despite lacking signals from particular communities (Crawford 2013) or interpretive frameworks. For instance, blog posts or tweets can be analysed out of the context in which they are posted (Boyd and Crawford 2012, p. 672), such as responses to a news item or as part of a sarcastic dialogue. This may in part be a technical limitation: the 'sensitivity' of social media data gathered via APIs is largely unclear at the time of collection for researchers, meaning seemingly innocuous 'chatter' can, when connected to other pieces of data, become highly sensitive and revealing about the data subject (Lomborg and Bechmann 2014, p. 261). The reviewed literature goes so far as to acknowledge the potentially problematic epistemic implications of such acts of interpretation, while stopping short of making a connection with normative or ethical implications.

4.2.5 The Big Data Divide

As with nearly any modern information and communication technology or practice, 'digital divides' can exist within Big Data practices. For example, individuals that 'opt out' of data collection may experience "exclusion from the digitally connected world in which they reside" (Nunan and Di Domenico 2013, p. 7). However, the term 'Big Data divide' is used to describe related but qualitatively different phenomena, understood as the inequalities between data subjects providing the material for Big Data analytics and the organisations with the necessary infrastructure and resources to analyse and understand the data (Andrejevic 2014; Crawford et al. 2014). A divide is created in the terms of the 'forms of knowing' made possible (Andrejevic 2014, p. 1676). The divide concerns 'haves' and 'have nots' of Big Data, where the ability and thus opportunities to assess and utilise Big Data are located within the few organisations possessing the required access, knowledge, computational, and organisational resources necessary to analyse and understand Big Datasets (Berry 2011, p. 2011; Boyd and Crawford 2012; Fairfield

and Shtein 2014, p. 46; McNeely and Hahm 2014, p. 308; Puschmann and Burgess 2014, p. 1694). Such a divide can already be seen for research via social media, where access to data from APIs is greatly restricted for individual researchers when compared to organisations or research groups that can pay for access (Lomborg and Bechmann 2014, p. 256; Schroeder 2014).

Big Data is increasingly becoming the sole domain of large organisations, despite calls to allow data subjects to benefit from and manipulate their data (Boyd and Crawford 2012; Tene and Polonetsky 2013). This situation can be troublesome for several reasons, foremost due to the inability of 'underprivileged' individual data subjects and organisations both to understand and have access to the methods, logic or at least "decisional criteria" behind Big Data analysis and decision-making processes (Tene and Polonetsky 2013, p. 243). Furthermore, it is often unclear which individuals and organisations can access or buy one's data (McNeely and Hahm 2014, p. 308).

The divide can also be conceived in terms of access to modify the data (Boyd and Crawford 2012, p. 674), or whether data subjects are empowered to be notified when data about them are created, modified or analysed, and given fair opportunities to access the data and correct errors or misinterpretations in the data and knowledge and profiles built upon it (Coll 2014). Superficially, such potential 'rights' can be connected to the 'right to be forgotten'[13] (Higuchi 2013), insofar as similar rights to modify privately held personal data (rather than publicly available links) could conceivably be granted as an oversight mechanism. Hypothetically, a right to 'self-determination' can ground such connected data rights (Coll 2014, p. 1258) to combat the 'transparency asymmetry' that exists when consumers lack information about how data about them is "collected, analysed and used" (Coll 2014, p. 1259; Richards and King 2013). Broader social "inequalities and biases" can therefore have uninhibited influence over data analysis where subjects lack oversight (McNeely and Hahm 2014, p. 308; Oboler et al. 2012, p. 3).

Profiling and Surveillance

A lack of oversight means data subjects are unaware of the decisions made about their data, and the criteria and categories into which their data fit. Decisions made on the basis of Big Data in some way may restrict the treatment, information or opportunities offered to data subjects (Tene and Polonetsky 2013, p. 252). These decisions made on the basis of aggregated data affect the individual behind the (de-identified) profile as a member of a group or category; "the profile and the person intersect" (Andrejevic 2014, p. 1677) quite apart from the individual's identity. Understanding when and why one's data have been 'categorised' as a particular type or instance of a particular phenomenon is therefore key to reinforcing self-

[13]For further details on the specification of the right to be forgotten by Google in the EU, see: Advisory Council to Google on the Right to be Forgotten (2015).

control of data and reducing the imbalance of power characteristic of the 'Big Data divide' (Lyon 2003). The 'data poor' are caught in a position of weakness wherein the ability to understand the data and methods used to make decisions about them as individuals and members of groups is beyond their means (Andrejevic 2014, p. 1678). Even where discrimination does not occur, "the relegation of decisions about an individual's life to automated processes" (Tene and Polonetsky 2013, p. 252) is itself troubling due to the imbalance in knowledge and decision-making power inherent in this setup.

Lupton (2014) describes this phenomenon in terms of analytic metrics used to sort individuals and groups and highlight specific aspects or characteristics to 'understand' them. The implicit interpretation behind supposedly 'objective' Big Data analysis can be seen in these metrics used within aggregated datasets. Metrics "make visible aspects of individuals and groups that are not otherwise perceptible, because they are able to join-up a vast range of details derived from diverse sources" (Lupton 2014, p. 859). These metrics provide different ways of 'seeing' the groups and interpreting their behaviours; whether a particular interpretation is correct or reflective of the meaning, identities or motivations given to acts by members of the group is unclear. Following on from the inability to modify or correct one's data (see Sect. 4.2.5), a 'right to be forgotten' according to which data subjects can request deletion or correction of particular pieces of data is thought to be more empowering and privacy-protecting than a blanket right to have a person's profile or entire data set deleted (Oboler et al. 2012, p. 9). Correcting the underlying data means future metrics will ideally be applied to a more 'accurate' or representative picture of the data subject in her terms.

Profiling can quickly take on surveillance implications (Bonilla 2014, p. 265); Big Data has been compared to an omniscient 'transparent human' capable of mass surveillance (Markowetz et al. 2014, p. 410). However, profiling need not be seen as a surveillance practice for concerns over profiling to be relevant—it is the act of interpreting the data through a particular framework of understanding or metric to 'make sense' of it, rather than any (problematic) actions taken once this sorting has occurred, which constitutes profiling.

Once profiled, actions taken towards particular groups may be problematic. To take an example from biomedicine, the extent to which data subjects are informed about research results, such as disease proclivity, may require new policies of professional conduct concerning when and how results are released to data subjects sorted into particular disease groups (McGuire et al. 2008, p. 1862, 2012).[14] Discrimination and benefits of Big Data may become localised around groups that present easy or interesting analysis opportunities. Crawford et al. (2014, p. 1667) argue that Big Data leads to new concentrations of power, 'blind spots' and problems

[14]Regulatory action may be required, as Big Data creates new opportunities for "data aggregators and miners to . . . run around health care's domain-specific protections by creating medical profiles of individuals" not subject to existing legislation (Terry 2012, p. 386), as was the case with the Google Health platform which operated outside of HIPAA restrictions in the United States (Mora 2012, p. 373).

of representativeness because it "cannot account for those who participate in the social world in ways that do not register as digital signals." Correcting these gaps is unlikely, as "big data's opacity to outsiders and subsequent claims to veracity through volume . . . discursively neutralizes the tendency to make errors." These 'blind spots' mean that analysis will tend to focus on data subjects and phenomena amenable to digitisation and measurement, meaning that the benefits and ethical burdens of Big Data will be placed, for better or worse, on specific social, cultural and economic groups (Majumder 2005, p. 37; McGuire et al. 2008). For instance, analysis of social media datasets will necessarily affect social media users and their underlying demographics in the first instance.

Justice

It may be possible to express such divides as ethically problematic in terms of justice. Interventions and knowledge developed from Big Data, particularly genomic and microbiomic data (Lewis et al. 2012), may favour populations from whom data is collected, further exacerbating existing gaps in medical practice and knowledge between "Euro-Americans of middle to upper socio-economic status" and others (Lewis et al. 2012, p. 2). Even where studied populations are diverse, formal benefit sharing agreements may be required between data subjects and custodians or researchers to ensure data are not taken from one context purely to benefit individuals in another, similar to the issues faced with pharmaceutical research in the third world (Mathaiyan et al. 2013, p. 103). As much should be done to facilitate benefit sharing as possible (Choudhury et al. 2014, p. 4), as Big Data can allow researchers to meet the moral obligation to maximise the value of data collected from research participants without the need for further data collection which places participants at risk (Currie 2013; Mello et al. 2013, p. 1653).

5 Discussion

Reviewing literature is a first step to conduct ethical foresight, in the sense that it allows one to distinguish between issues and implications that are currently under consideration, and those that are not yet acknowledged or require further attention. Overall, the quality of the reviewed literature leaves gaps based on a dearth of empirical research and 'deep' conceptual analysis. In particular, the prevalence of 'opinion pieces' and 'editorials' that briefly raise issues but do not discuss them in depth shows the need for further scholarship in this area of emerging ethical import.

As the results were presented as a narrative overview with accompanying commentary, this section will take the next step by drawing attention to issues that have received insufficient attention in the literature. Specifically, the discussion highlights issues that are expected by the authors to be key ethical issues in the near future, and which require further exploration in the context of specific Big Data

practices and domains. These issues include group-level ethics, ethical implications of growing epistemological challenges (e.g. Floridi 2012), effects of Big Data on fiduciary relationships, the ethics of academic versus commercial practices, ownership of intellectual property derived from Big Data, and the content of and barriers to meaningful data access rights.

5.1 Group-Level Ethics

Technological means to prevent ethical problems through Big Data tend to focus on the individual, ignoring harms which affect groups. Data protection legislation and anonymisation techniques implicitly focus on the individual in seeking an appropriate balance between the value of the anonymised dataset for subsequent analysis and the privacy of individual data subjects. Such technical solutions to avoid the potential ethical harms of Big Data practices are only partially successful and remain fallible. Advances in analytic methods and technologies of re-engineering identity (e.g. Cassa et al. 2008; Hay et al. 2008), or failures in the oversight processes preceding the release of datasets which fail to identify potential means of re-identification guarantee future vulnerability.

In the face of such technological and practical uncertainties (e.g. Mittelstadt et al. 2015), employing punitive measures for attempts to re-identify data, or emphasising professional responsibility (for example through codes of ethics for data custodians; see Sect. 5.3 and Oboler et al. 2012, p. 11) may prove more effective than increasingly restrictive anonymisation protocols. Alternatively, data may be hosted in 'safe harbours' within which data uses are screened and controlled (Dove et al. 2014). Although these measures do not address group-level effects, they are pragmatically responsive to possibilities of re-identification, while not further restricting movement of anonymised data.

Even where such solutions are implemented, the emphasis on protecting the individual problematically focuses ethical assessment on harms at the individual level (see section "Anonymisation"); perfectly anonymised datasets still allow for group-level ethical harms for which the identities of members of the group or profile are irrelevant (Sloot 2014). Algorithmic grouping of data points and identification of statistical relationships allows for profiling and grouping of individual data subjects (see section "Profiling and Surveillance"). Profiling connects data subjects to one another, meaning the behaviours, preferences and interests of others affect how the individual is treated in ethically relevant ways. Preferential treatment and decision-making in a variety of contexts of variable ethical acceptability can be justified on this basis, such as personalised pricing in e-commerce or genetic discrimination.[15]

[15] As an example of the latter, if biobanking research utilising genome sequences were to reveal that obesity is linked primarily to behaviour rather than genes, or an ethnic group were shown to have

To address potential discrimination against particular demographic, genomic or other groups, an 'ethic of care' approach may be required which would set aside particular forms of research or hypotheses as 'off limits' (cf. Lewis et al. 2012). Alternatively, it may be possible to conceive of privacy as a group-level concept and thus speak of 'group privacy rights' that could restrict the flow and acceptable uses of aggregated datasets and profiling. However, the feasibility and practicalities of expanding privacy rights require further investigation, in particular the potential barriers created for desirable research similar to the informed consent debate currently underway in Europe (see Sect. 4.2.1; Taylor and Floridi 2016).

5.2 Epistemological Difficulties

As discussed above, a loss of qualification or contextual aspects of data has been observed in Big Data analytics, which in some cases can be attributed to complex interpretations of data performed by computers or analytical algorithms (Bowker 2013, p. 170). While this position problematically appears to place the responsibility for interpretation (seeing data *as* something) entirely on (learning) algorithms while exonerating designers of algorithms and the ontological categories within which they interpret, it helpfully emphasises the loss of context through quantification and categorisation of diverse datasets to facilitate analysis and connectivity. This loss of context or 'decontextualisation' can be understood as an instance of 'ontic occlusion' (see Bowker 2014; Knobel 2010), or the process by which emphasising particular aspects of a phenomenon in a discourse necessarily occludes or 'downplays' other aspects.

Ontic occlusion, originally developed to describe ontological characteristics of archiving, can be extended to Big Data to describe a qualitative loss or degradation of the data implied by acts of interpretation, classification or categorisation of the data in collection and analysis. Archives or datasets, conceived of as discourses, "cannot in principle contain the world in small . . . most slices of reality are not represented" (Bowker 2014, p. 1797). If data is seen as describing a particular instance of a phenomenon, for example data describing the case of a particular cancer patient, the instance and data become equivalent; the profile becomes a representation of the profiled (e.g. Floridi 2012). While undoubtedly a problem with any type of data collection and analysis, in Big Data this necessary loss of context is exacerbated by the sheer scale of data being analysed. It is tempting to view the profile, or the data, as representative of the whole phenomenon (Bowker 2014, p. 1797); increasing the scale of data to be considered only increases the difficulty of identifying what is stripped from data to make sense of it. The implications of this problem require further attention in specific Big Data practices; for example,

a higher genetic pre-disposition to cancer (cf. Angrist 2009; Mathaiyan et al. 2013), well-meaning research may inadvertently lead to future discrimination against these groups.

it is likely more ethically problematic to strip context from data used to track the behaviours of individuals than it is to remove identifying information from tissue samples for medical research.

5.3 Fiduciary Relationships

Further research may also be required into the effects of Big Data on the 'internal goods' (cf. MacIntyre 2007) of relationships and interactions between data custodians (e.g. researchers, commercial organisations, repositories) and data subjects. The background disciplines and sentiments informing the conceptualisation of 'Big Data' in ongoing discussion is important in defining the obligations that can be attributed to data custodians. When Big Data is thought of as a form of business based around the selling and processing of data for commercial advantage, it is perhaps inappropriate to expect a relationship based on 'trust' or professionalism to exist between subjects and custodians (cf. Terry 2012).

The mediating role of data in these relationships, by which data subjects are 'represented' or revealed to custodians through their data, may be of ethical importance in certain contexts. In medicine for example, greater reliance on data representations of patients brought about by adoption of Big Data practices may create new gaps in care or doctor-patient relationships (cf. Beauchamp and Childress 2009; MacIntyre 2007; Pellegrino and Thomasma 1993). Traditional fiduciary 'healing relationships' do not scale well to Big Data or even institutional care (Terry 2014, p. 838), meaning that, as data representations and models are increasingly used to understand the patient's condition, the 'virtues' or internal goods of traditional medical relationships may be subtly undermined or realised less frequently.

Harm can occur to the data subject through misinterpretation or overreliance on data representing the subject's state(s) of being. The 'goods' provided by such relationships, which extend beyond issues of efficiency or effectiveness of interventions and are derived from the character of the individual providing care, may be undermined; for instance, care providers may be less able to demonstrate understanding, compassion and other desirable traits found within 'good' medical interactions *in addition to* applying their knowledge of medicine to the patient's case (cf. Beauchamp and Childress 2009; MacIntyre 2007; Pellegrino and Thomasma 1993). Put another way, the patient's body and voice may increasingly be replaced or supplemented by data representations of state of being if Big Data practices are adopted in medicine (Barry et al. 2001). Further research is required into the effects of these representations on the quality of relationships through which care is provided. Medical relationships are of particular concern due to the patient being in a vulnerable (and trusting) state (Pellegrino and Thomasma 1993).

5.4 Academic vs. Commercial Practices

In terms of the likelihood of future problematic uses, a distinction should be drawn between 'academic' and 'commercial' Big Data practices in order to allow data subjects to retain realistic expectations over potential uses and implications of authoring data (cf. Lupton 2014). The need for such a distinction can be seen for example in the deficiencies of existing patient experience websites, many of which fail to inform users whether collected data will be used for research or commercial purposes (Lupton 2014), or in ethically controversial research being permitted in commercial contexts which would not pass the scrutiny of an academic ethical review board (Schroeder 2014). While 'research' and 'commercialisation' are not mutually exclusive, meaningful ethical distinctions can be drawn. The purpose here is not to distinguish between types of Big Data practices, but rather the motivations behind them. For example, commercial and academic research may be qualitatively similar, in terms of the experiences of the data subject and methods of research, but differ substantially in motives, e.g. basic research to advance scientific knowledge versus product development. Furthermore, data subjects may be interested in the degree of oversight for particular practices. In general, research-based practices will require some form of ethical review and monitoring, whereas commercial practices will not. Clearly, this distinction requires further specification to distinguish between 'types' of Big Data practices in terms of their ethical dimensions.

5.5 Ownership of Intellectual Property

In the reviewed literature, ownership was discussed as a mechanism to control data. While undoubtedly important, ownership can also refer to owning products and intellectual property produced through Big Data practices. This issue was only discussed in one article which called for benefit sharing with data subjects to allow for innovation led by data subjects in developing products and services from Big Data (Tene and Polonetsky 2013). Despite this relative paucity of attention, this topic deserves further debate due to the potential to develop commercially valuable material through analysis of data collected from or volunteered by members of the public. Currently, data subjects tend not to benefit from analysis of data collected about them—users of Facebook, for instance, do not share in the revenue derived from targeted advertisements. As similar products and services become increasingly common and commercially viable in the future, the ownership of personal data will attain renewed importance. In the future, Big Data will likely raise questions over ownership structures in which data subjects forfeit all rights to personal data generated through usage of networked products and services. It could alternatively become the norm for data subjects to share in (financial) benefits derived from their data, or at least to be guaranteed access to it for personal uses and development. At the very least, ownership structures for personal data require further attention due to the apparent potential of Big Data, to encourage and exploit exponential growth of personal data.

5.6 Data Access Rights

Following on from ownership, access mechanisms and rights for data subjects require further attention. As discussed in the context of ownership (see Sect. 4.2.3), data subject rights to access and modify data are reliant upon the subject being aware of what data exist about her, who holds them, what they (potentially) mean and how they are being used. Assuming such rights are sought (as specified in data protection legislation, for example), significant technical and practical barriers to their realisation exist which may be insurmountable, thus precluding the possibility of meaningful data access rights in the era of Big Data.

For access rights to be meaningful, data subjects must be able to exercise them with reasonable effort. For instance, being provided with thousands of printed pages of digital data would require unreasonable effort on the part of the data subject to compile and understand the data, and would therefore fail to preserve a meaningful right to access. As discussed in the context of the 'Big Data divide', resource, skillset and comprehension barriers exist which would prevent a 'lay' data subject from being able to exercise the aforementioned access rights. Big Data requires significant computational power and storage, and advanced scientific know-how. As with any data science, analysis will require discipline-specific skills and knowledge, often only accessible through extensive training and education. Even for willing subjects, the amount of time and effort required to attain the background knowledge and skills to understand the totality of data held about oneself may easily be overwhelming. Ascertaining the extent and uses of data held about an individual is also difficult, given the often 'hidden' and seemingly ubiquitous nature of personal data processing (see Sect. 2).

Considered together, the emerging picture is of data subjects in a disempowered state, faced with seemingly insurmountable barriers to understanding *who* holds *what* data about them, being used for *which* purposes. Further, in relation to modification and correction of personal data, it is unclear how subjects can possibly propose changes to data without first understanding the contents and inferences drawn from them, or the perhaps inaccurate or incomplete ways in which the data represent the subject and her behaviours. For a meaningful right to modification and correction it may therefore be necessary for data custodians to provide oversight and explanations of categories, profiles or other criteria used in sorting the data to, at a minimum, allow subjects to understand the 'silos' into which they have been placed (see section "Profiling and Surveillance").

Considered together, these barriers may preclude the exercise of meaningful data access rights within current Big Data practices. However, further research is required to justify this assertion. Specifically, specifications are required of reasonable access rights, domain-specific barriers to access, and alterations to practices or data protection legislation which will ensure data custodians assist data subjects in gaining meaningful access as far as possible.

A small number of mechanisms to address issues of data sharing and irresponsible usage of data have been proposed in the reviewed literature. For instance,

McNeely and Hahm (2014, p. 1654) have proposed a set of 'core principles of expanded data sharing' to be followed by "any system that is ultimately adopted for expanded access to participant-level data." These principles emphasise responsibility, privacy, equal treatment of all data requesters/trial sponsors, accountability of data custodians and requesters, and the practicality of the system in terms of transparent and timely responses to data requests and a lack of other such unnecessary barriers to access. Other suggestions include granting data subjects a 'right to be forgotten', a 'right to data expiry', and the 'ownership of a social graph'. The first refers to the ability of data subjects to request that links to information about them be deleted. The second refers to the automatic deletion of unstructured data after a set period of time if they no longer have any commercial or research value. The third will detail what data exist about an individual, when and how they were collected, and where they are stored (Nunan and Di Domenico 2013).

While each of these concepts faces theoretical and practical difficulties, such as defining 'commercial' or 'research' value, they nevertheless represent an attempt to realise meaningful data rights in the era of Big Data. Modifications appear to be required given the existing inaccessibility and incomprehensibility of Big Data algorithms and practices to 'lay' data subjects—some form of assistance or 'hand holding' is required by data custodians given the increasing prevalence of data in mediating human interactions. Going forward, competitive interests and desires for commercial secrecy need to be balanced against meaningful access rights for data subjects.

6 Conclusion

As is often the case with emerging technologies and sciences, a tendency has been recognised to overemphasise the potential benefits of Big Data as a means of explaining 'everything', perhaps without the need for theories or frameworks of understanding (Callebaut 2012; Crawford 2013). "Data fundamentalism," or the idea that "correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth" (Crawford 2013), problematically influences the public, mass media and researchers where a tendency exists to view the advancement of Big Data into all information-based disciplines as inevitable. In such cases, beneficial outcomes of this shift are often similarly 'inevitable' (e.g. Costa 2014, p. 436), with practitioners more concerned with communicating how 'good' or 'responsible' they are rather than investigating what these concepts mean in the context of specific Big Data practices. Such broad brush attitudes towards Big Data should be avoided if its ethical implications are to be given serious consideration throughout the life of emerging Big Data practices, products and applications.

The analysis offered in this article is intended to contribute to transforming such general and perhaps overly optimistic attitudes by providing a starting point and comprehensive reference for future discussions of the ethics of Big Data, especially

in the very sensitive context of biomedical research. An overview of key ethical issues of Big Data has been offered, against which areas requiring further research in the near term have been identified. In particular, biomedical applications of Big Data have been identified as particularly ethically challenging due to the sensitivity of health data and fiduciary nature of healthcare. It is our hope that the analysis will contribute to ethically responsibility development, deployment and maintenance of novel datasets and practices in biomedicine and beyond in the era of Big Data.

Acknowledgements The research leading to this work has been funded by a John Fell Fund major research grant. An initial version of this paper was discussed at a workshop organised at the Ethics of Biomedical Big Data workshop organised in April 2015 at the Oxford Internet Institute. We wish to acknowledge the extremely valuable feedback received during that meeting and from the two anonymous reviewers.

References

Advisory Council to Google on the Right to be Forgotten. 2015. Report of the Council to Google on the Right to be Forgotten. Google Docs. https://drive.google.com/file/d/0B1UgZshetMd4cEI3SjlvV0hNbDA/view?pli=1&usp=embed_facebook. Accessed 19 Mar 2015.

Andrejevic, M. 2014. Big data, big questions the big data divide. *International Journal of Communication* 8: 17. Accessed 7 October 2014.

Angrist, M. 2009. Eyes wide open: The personal genome project, citizen science and veracity in informed consent. *Personalized Medicine* 6: 691–699.

Apple. 2014. iBeacon for Developers – Apple Developer. https://developer.apple.com/ibeacon/. Accessed 17 Nov 2014.

Bail, C.A. 2014. The cultural environment: Measuring culture with big data. *Theory and Society* 43(3–4): 465–482. doi:10.1007/s11186-014-9216-5.

Barry, C.A., F.A. Stevenson, N. Britten, N. Barber, and C.P. Bradley. 2001. Giving voice to the lifeworld. More humane, more effective medical care? A qualitative study of doctor-patient communication in general practice. *Social Science and Medicine* 53: 487–505. doi:10.1016/s0277-9536(00)00351-8.

Beauchamp, T.L., and J.F. Childress. 2009. *Principles of biomedical ethics*. New York: Oxford University Press.

Berry, D. M. 2011. The computational turn: Thinking about the digital humanities. *Culture Machine* 12(0). ftp://121.171.90.140/big.data/%EB%B9%85%EB%8D%B0%EC%9D%B4%ED%84%B02_20131024_sunup/THE%20COMPUTATIONAL%20TURN%20Digital-Humanities.pdf. Accessed 7 Oct 2014.

Bonilla, D.N. 2014. Information management professionals working for intelligence organizations: Ethics and deontology implications. *Security and Human Rights* 24(3–4): 264–279. doi:10.1163/18750230-02404005.

Bowker, G. C. 2013. *Data flakes: An afterword to "Raw Data" is an oxymoron*. "Raw data" is an oxymoron. Cambridge, MA: MIT Press. http://www.ics.uci.edu/~vid/Readings/bowker_data_flakes.pdf. Accessed 14 Oct 2014.

Bowker, G.C. 2014. Big data, big questions the theory/data thing. *International Journal of Communication* 8: 5. Accessed 7 October 2014.

Boyd, danah., and K. Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society* 15(5): 662–679. doi:10.1080/1369118X.2012.678878.

Boye, N. 2012. Co-production of health enabled by next generation personal health systems. *Studies in Health Technology and Informatics* 177: 52–58.

Busch, L. 2014. Big data, big questions a dozen ways to get lost in translation: Inherent challenges in large scale data sets. *International Journal of Communication* 8: 18. Accessed 7 October 2014.

Butler, D. 2013. When Google got flu wrong. *Nature* 494(7436): 155–156. doi:10.1038/494155a.

Callebaut, W. 2012. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 69–80. doi:10.1016/j.shpsc.2011.10.007.

Cassa, C.A., S.C. Wieland, and K.D. Mandl. 2008. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics* 7(1): 45. doi:10.1186/1476-072X-7-45.

Choudhury, S., J.R. Fishman, M.L. McGowan, and E.T. Juengst. 2014. Big data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience* 8: 239. doi:10.3389/fnhum.2014.00239.

Clayton, E.W. 2005. Informed consent and biobanks. *Journal of Law, Medicine & Ethics* 33(1): 15–21. doi:10.1111/j.1748-720X.2005.tb00206.x.

Collingridge, D. 1980. *The social control of technology*. New York: Palgrave Macmillan.

Coll, S. 2014. Power, knowledge, and the subjects of privacy: understanding privacy as the ally of surveillance. *Information Communication & Society* 17(10): 1250–1263. doi:10.1080/1369118X.2014.918636.

Costa, F.F. 2014. Big data in biomedicine. *Drug Discovery Today* 19(4): 433–440. doi:10.1016/j.drudis.2013.10.012.

Craig, T. 2011. *Privacy and big data*. Sebastopol/Cambridge: O'Reilly.

Crawford, K. 2013. The hidden biases in big data. *Harvard Business Review*. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/. Accessed 10 Oct 2014.

Crawford, K., M.L. Gray, and K. Miltner. 2014. Critiquing big data: Politics, ethics, epistemology special section introduction. *International Journal of Communication* 8: 10. Accessed 2 October 2014.

Currie, J. 2013. "Big Data" versus "Big Brother": On the appropriate use of large-scale data collections in pediatrics. *The Journal of Pediatrics* 131(Suppl): S127–S132. doi:10.1542/peds.2013-0252c.

Dereli, T., Y. Coskun, E. Kolker, O. Guner, M. Agirbasli, and V. Ozdemir. 2014. Big data and ethics review for health systems research in LMICs: Understanding risk, uncertainty and ignorance-and catching the black swans? *American Journal of Bioethics* 14(2): 48–50. doi:10.1080/15265161.2013.868955.

Devos, Y., P. Maeseele, D. Reheul, L. Van Speybroeck, and D. De Waele. 2008. Ethics in the societal debate on genetically modified organisms: A (Re)Quest for sense and sensibility. *Journal of Agricultural and Environmental Ethics* 21(1): 29–61. doi:10.1007/s10806-007-9057-6.

Docherty, A. 2014. Big data – ethical perspectives. *Anaesthesia* 69(4): 390–391. doi:10.1111/anae.12656.

Dove, E.S., B.M. Knoppers, and M.H. Zawati. 2014. Towards an ethics safe harbor for global biomedical research. *Journal of Law and the Biosciences* 1(1): 3–51. doi:10.1093/jlb/lst002.

Enjolras, B. 2014. Big data and social research: New possibilities and ethical challenges. *Tidsskrift for Samfunnsforskning* 55(1): 80–89.

EURORDIS. 2013. Statement on the EP report on the protection of personal data. http://www.publichealth.ox.ac.uk/helex/Statement%20Data%20Prot%20FINAL.pdf. Accessed 22 Oct 2014.

Fairfield, J., and H. Shtein. 2014. Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics* 29(1): 38–51. doi:10.1080/08900523.2014.863126.

Fan, W., and A. Bifet. 2013. Mining big data: Current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter* 14(2): 1–5. Accessed 2 October 2014.

Floridi, L. 2008. The method of levels of abstraction. *Minds and Machines* 18(3): 303–329. doi:10.1007/s11023-008-9113-7.

Floridi, L. 2012. Big data and their epistemological challenge. *Philosophy & Technology* 25(4): 435–437. doi:10.1007/s13347-012-0093-4.

Floridi, L. 2013. *The philosophy of information*. Reprint edn. Oxford: OUP Oxford.

Floridi, L, ed. 2014a. *The onlife manifesto*. New York: Springer. http://www.springer.com/philosophy/epistemology+and+philosophy+of+science/book/978-3-319-04092-9. Accessed 2 Dec 2014.

Floridi, L. 2014b. Open data, data protection, and group privacy. *Philosophy & Technology* 27(1): 1–3. doi:10.1007/s13347-014-0157-8.

Gadamer, H.G. 1976. *The historicity of understanding*. Harmondsworth: Penguin Books Ltd.

Gadamer, H.G. 2004. *Truth and method*. London: Continuum International Publishing Group.

General Medical Council. 2008. Consent guidance. http://www.gmc-uk.org/guidance/ethical_guidance/consent_guidance_index.asp

Gilligan, C. 1982. *In a different voice*. Cambridge: Harvard University Press.

Goodman, E. 2014. Design and ethics in the era of big data. *Interactions* 21(3): 22–24. Accessed 1 October 2014.

Habermas, J. 1984. *The theory of communicative action: Volume 1: Reason and the rationalization of society*. Boston: Beacon.

Habermas, J. 1985. *The theory of communicative action: Volume 2: Lifeworld and system: A critique of functionalist reason*. Boston: Beacon.

Hansson, M.G. 2009. Ethics and biobanks. *British Journal of Cancer* 100(1): 8–12. doi:10.1038/sj.bjc.6604795.

Harris, J. 2005. Scientific research is a moral duty. *Journal of Medical Ethics* 31(4): 242–248. doi:10.1136/jme.2005.011973.

Hayden, E. C. 2012. *A broken contract*. London: Nature Publishing Group Macmillan Building. http://environmentportal.in/files/file/informed%20consent.pdf. Accessed 7 Oct 2014.

Hay, M., G. Miklau, D. Jensen, D. Towsley, and P. Weis. 2008. Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment* 1(1): 102–114. doi:10.14778/1453856.1453873.

Heidegger, M. 1967. *Being and time*. Malden: Blackwell.

Helbing, D., and S. Balietti. 2011. From social data mining to forecasting socio-economic crises. *European Physical Journal-Special Topics* 195(1): 3–68. doi:10.1140/epjst/e2011-01401-8.

Higuchi, N. 2013. Three challenges in advanced medicine. *Japan Medical Association Journal* 56(6): 437–447.

Hoffman, S. 2014. *Citizen science: The law and ethics of public access to medical big data* (SSRN Scholarly Paper No. ID 2491054). Rochester: Social Science Research Network. http://papers.ssrn.com/abstract=2491054. Accessed 13 Oct 2014.

Hoffman, S., and A. Podgurski. 2013. Big bad data: Law, public health, and biomedical databases. *Journal of Law, Medicine and Ethics* 41(SUPPL. 1): 56–60. doi:10.1111/jlme.12040.

IBM. 2014. The four V's of Big Data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data. Accessed 23 Oct 2014.

Ioannidis, J.P.A. 2013. Informed consent, big data, and the oxymoron of research that is not research. *American Journal of Bioethics* 13(4): 40–42. doi:10.1080/15265161.2013.768864.

Joly, Y., E.S. Dove, B.M. Knoppers, M. Bobrow, and D. Chalmers. 2012. Data sharing in the post-genomic world: The experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Computational Biology* 8(7), e1002549. doi:10.1371/journal.pcbi.1002549.

Kass, N.E. 2001. An ethics framework for public health. *American Journal of Public Health* 91(11): 1776–1782. doi:10.2105/AJPH.91.11.1776.

Kaye, J., L. Curren, N. Anderson, K. Edwards, S.M. Fullerton, N. Kanellopoulou, et al. 2012. From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews Genetics* 13(5): 371–376. doi:10.1038/nrg3218.

Knobel, C. P. 2010. Ontic occlusion and exposure in sociotechnical systems. University of Pittsburgh. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/78763

Krotoski, A.K. 2012. Data-driven research: Open data opportunities for growing knowledge, and ethical issues that arise. *Insights: the UKSG Journal* 25(1): 28–32. doi:10.1629/2048-7754.25.1.28.

Laney, D. 2001. 3D data management: Controlling data volume, velocity and variety. META Group Research Note 6.

Larson, E.B. 2013. Building trust in the power of "big data" research to serve the public good. *JAMA Journal of the American Medical Association* 309(23): 2443–2444. doi:10.1001/jama.2013.5914.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, et al. 2009. Computational social science. *Science* 323(5915): 721–723. doi:10.1126/science.1167742.

Lewis, C.M., A. Obregón-Tito, R.Y. Tito, M.W. Foster, and P.G. Spicer. 2012. The human microbiome project: Lessons from human genomics. *Trends in Microbiology* 20(1): 1–4. doi:10.1016/j.tim.2011.10.004.

Liyanage, H., S. de Lusignan, S.-T. Liaw, C.E. Kuziemsky, F. Mold, P. Krause, et al. 2014. Big data usage patterns in the health care domain: A use case driven approach applied to the assessment of vaccination benefits and risks. Contribution of the IMIA primary healthcare working group. *Yearbook of Medical Informatics* 9(1): 27–35. doi:10.15265/IY-2014-0016.

Lomborg, S., and A. Bechmann. 2014. Using APIs for data collection on social media. *Information Society* 30(4): 256–265. doi:10.1080/01972243.2014.915276.

Lupton, D. 2014. The commodification of patient opinion: The digital patient experience economy in the age of big data. *Sociology of Health & Illness* 36(6): 856–869. doi:10.1111/1467-9566.12109.

Lyon, D. 2003. *Surveillance as social sorting : Privacy, risk, and digital discrimination*. London: Routledge.

MacIntyre, A. 2007. *After virtue: A study in moral theory*, 3rd ed. London: Gerald Duckworth & Co Ltd.

Mahajan, R. L., Reed, J., Ramakrishnan, N., Mueller, R., Williams, C. B., and Campbell, T. A. 2012. Cultivating emerging and black swan technologies. *Presented at the ASME International Mechanical Engineering Congress and Exposition, Proceedings (IMECE)* 6: 549–557. doi:10.1115/IMECE2012-89339.

Majumder, M.A. 2005. Cyberbanks and other virtual research repositories. *Journal of Law, Medicine & Ethics* 33(1): 31–39. doi:10.1111/j.1748-720X.2005.tb00208.x.

Markowetz, A., K. Błaszkiewicz, C. Montag, C. Switala, and T.E. Schlaepfer. 2014. Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses* 82(4): 405–411. doi:10.1016/j.mehy.2013.11.030.

Master, Z., L. Campo-Engelstein, and T. Caulfield. 2014. Scientists' perspectives on consent in the context of biobanking research. *European Journal of Human Genetics* 23(5): 569–574. doi:10.1038/ejhg.2014.143.

Mathaiyan, J., A. Chandrasekaran, and S. Davis. 2013. Ethics of genomic research. *Perspectives in Clinical Research* 4(1): 100. doi:10.4103/2229-3485.106405.

McGuire, A.L., L.S. Achenbaum, S.N. Whitney, M.J. Slashinski, J. Versalovic, W.A. Keitel, and S.A. McCurdy. 2012. Perspectives on human microbiome research ethics. *Journal of Empirical Research on Human Research Ethics: An International Journal* 7(3): 1–14. doi:10.1525/jer.2012.7.3.1.

McGuire, A.L., J. Colgrove, S.N. Whitney, C.M. Diaz, D. Bustillos, and J. Versalovic. 2008. Ethical, legal, and social considerations in conducting the human microbiome project. *Genome Research* 18(12): 1861–1864. doi:10.1101/gr.081653.108.

McNeely, C.L., and J. Hahm. 2014. The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research* 31(4): 304–310. doi:10.1111/ropr.12082.

Mello, M.M., J.K. Francer, M. Wilenzick, P. Teden, B.E. Bierer, and M. Barnes. 2013. Preparing for responsible sharing of clinical trial data. *New England Journal of Medicine* 369(17): 1651–1658. doi:10.1056/NEJMhle1309073.

Mittelstadt, B. D., Fairweather, N. B., McBride, N., and Shaw, M. 2011. Ethical issues of personal health monitoring: A literature review. In *ETHICOMP 2011 conference proceedings*, 313–321. Presented at the ETHICOMP 2011, Sheffield.

Mittelstadt, B. D., Fairweather, N. B., McBride, N., and Shaw, M. 2013. Privacy, risk and personal health monitoring. In *ETHICOMP 2013 conference proceedings*, 340–351. Presented at the ETHICOMP 2013. Kolding.

Mittelstadt, B.D., N.B. Fairweather, M. Shaw, and N. McBride. 2014. The ethical implications of personal health monitoring. *International Journal of Technoethics* 5(2): 37–60.

Mittelstadt, B.D., B.C. Stahl, and N.B. Fairweather. 2015. How to shape a better future? Epistemic difficulties for ethical assessment and anticipatory governance of emerging technologies. *Ethical Theory and Moral Practice* 18(5): 1027–1047.

Moore, P., Xhafa, F., Barolli, L., and Thomas, A. 2013. Monitoring and detection of agitation in dementia towards real-time and big-data solutions. 2013 Eighth international conference on P2p, parallel, grid, cloud and internet computing (3pgcic 2013), 128–135. doi:10.1109/3PGCIC.2013.26.

Moor, J. 1985. What is computer ethics?*. *Metaphilosophy* 16(4): 266–275. doi:10.1111/j.1467-9973.1985.tb00173.x.

Mora, F. 2012. The demise of Google health and the future of personal health records. *International Journal of Healthcare Technology and Management* 13(5): 363–377. Accessed 11 November 2014.

National Science Foundation. 2014. Critical techniques and technologies for advancing big data science & engineer (BIGDATA) – Program solicitation NSF 14-543. http://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf. Accessed 17 Oct 2014.

NHS England. (2014). NHS England: The care.data programme – Better information means better care. http://www.england.nhs.uk/ourwork/tsd/care-data/. Accessed 11 Nov 2014.

Niemeijer, A.R., B.J. Frederiks, I.I. Riphagen, J. Legemaate, J.A. Eefsting, and C.M. Hertogh. 2010. Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: An overview of the literature. *International Psychogeriatrics* 22: 1129–1142.

Nissenbaum, H. 2004. *Privacy as contextual integrity* (SSRN scholarly paper no. ID 534622). Rochester: Social Science Research Network. http://papers.ssrn.com/abstract=534622. Accessed 12 Mar 2013.

Noddings, N. 2013. *Caring: A relational approach to ethics and moral education*. Berkeley: Univ of California Press.

Nuffield Council on Bioethics. 2015. *The collection, linking and use of data in biomedical research and health care: ethical issues*, 198. Nuffield Council on Bioethics. http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf

Nunan, D., and M. Di Domenico. 2013. Market research and the ethics of big data. *International Journal of Market Research* 55(4): 505. doi:10.2501/IJMR-2013-015.

Oboler, A., Welsh, K., and Cruz, L. 2012. The danger of big data: Social media as computational social science. *First Monday* 17(7). https://www.scopus.com/inward/record.url?eid=2-s2.0-84867308941&partnerID=40&md5=0e4cb2f657154c7f82a76c2a657259ab

Pariser, E. 2011. *The filter bubble : What the internet is hiding from you*. London: Viking.

Patterson, M. E., and Williams, D. R. 2002. *Collecting and analyzing qualitative data: Hermeneutic principles, methods and case examples*, Vol. 9. Champaign: Sagamore Publishing, Inc. http://www.treesearch.fs.fed.us/pubs/29421. Accessed 7 Nov 2012.

Pellegrino, E.D., and D.C. Thomasma. 1993. *The virtues in medical practice*. New York: Oxford University Press.

Prainsack, B., and A. Buyx. 2013. A solidarity-based approach to the governance of research biobanks. *Medical Law Review* 21(1): 71–91. doi:10.1093/medlaw/fws040.

Puschmann, C., and J. Burgess. 2014. Big data, big questions metaphors of big data. *International Journal of Communication* 8: 20. Accessed 7 October 2014.

Reuters. 2014, October 3. Facebook plots first steps into healthcare. http://www.telegraph.co.uk/technology/facebook/11139606/Facebook-plots-first-steps-into-healthcare.html. Accessed 15 Nov 2014.

Richards, N.M., and J.H. King. 2013. Three paradoxes of big data. *Stanford Law Review Online* 66: 41. Accessed 18 February 2015.

Rothstein, M.A., and A.B. Shoben. 2013. An unbiased response to the open peer commentaries on "does consent bias research?". *The American Journal of Bioethics* 13(4): W1–W4. doi:10.1080/15265161.2013.769824.

Safran, C., M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, et al. 2006. Toward a national framework for the secondary use of health data: An American medical informatics association white paper. *Journal of the American Medical Informatics Association* 14(1): 1–9. doi:10.1197/jamia.M2273.

Schadt, E.E. 2012. The changing privacy landscape in the era of big data. *Molecular Systems Biology* 8: 612. doi:10.1038/msb.2012.47.

Schaefer, G.O., E.J. Emanuel, and A. Wertheimer. 2009. The obligation to participate in biomedical research. *Journal of the American Medical Association* 302(1): 67–72. Accessed 19 March 2015.

Schroeder, R. 2014. Big Data and the brave new world of social media research. *Big Data & Society* 1(2). doi:10.1177/2053951714563194.

Schroeder, R., and Cowls, J. 2014. Big Data, ethics, and the social implications of knowledge production. http://dataethics.github.io/proceedings/BigDataEthicsandtheSocialImplicationsofKnowledgeProduction.pdf. Accessed 2 Oct 2014.

Schwandt, T.A. 2000. Three epistemological stances for qualitative inquiry: Interpretivism, hermeneutics, and social constructionism. In *Handbook of qualitative research*, 189–214. Thousand Oaks: Sage.

Shilton, K. 2012. Participatory personal data: An emerging research challenge for the information sciences. *Journal of the American Society for Information Science and Technology* 63(10): 1905–1915. doi:10.1002/asi.22655.

Sloot, B. V. der. 2014. Privacy in the post-NSA era: Time for a fundamental revision?. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432104. Accessed 17 Feb 2015.

Slote, M. 2007. *The ethics of care and empathy*, New Ed edn. London/New York: Routledge.

Steinsbekk, K.S., L.Ø. Ursin, J.-A. Skolbekken, and B. Solberg. 2013. We're not in it for the money—Lay people's moral intuitions on commercial use of "their" biobank. *Medicine, Health Care and Philosophy* 16(2): 151–162. doi:10.1007/s11019-011-9353-9.

Taylor, L., and L. Floridi (eds.). 2016 (in press). *Group privacy – New challenges of data technologies*. New York: Springer.

Tene, O., and Polonetsky, J. 2013. Big data for all: Privacy and user control in the age of analytics. http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20. Accessed 2 Oct 2014.

Terry, N. 2012. Protecting patient privacy in the age of big data. *UMKC Law Review* 81: 385. Accessed 2 October 2014.

Terry, N. 2014. Health privacy is difficult but not impossible in a post-hipaa data-driven world. *Chest* 146(3): 835–840. doi:10.1378/chest.13-2909.

The NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, et al. 2009. The NIH human microbiome project. *Genome Research* 19(12): 2317–2323. doi:10.1101/gr.096651.109.

Watson, R.W.G., E.W. Kay, and D. Smith. 2010. Integrating biobanks: Addressing the practical and ethical issues to deliver a valuable tool for cancer research. *Nature Reviews Cancer* 10(9): 646–651. doi:10.1038/nrc2913.

Wellcome Trust. 2013. Impact of the draft European data protection regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research. Wellcome Trust. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf. Accessed 22 Oct 2014.