Methods and Biostatistics in Oncology

Understanding Clinical Research as an Applied Tool

Raphael L. C. Araújo
Rachel P. Riechelmann
Editors

Springer

Methods and Biostatistics in Oncology

Raphael L. C. Araújo • Rachel P. Riechelmann
Editors

Methods and Biostatistics in Oncology

Understanding Clinical Research as an Applied Tool

Springer

Editors
Raphael L. C. Araújo
Hospital do Câncer de Barretos
Barretos, SP, Brazil

Rachel P. Riechelmann
Department of Clinical Oncology
AC Camargo Cancer Center
São Paulo, SP, Brazil

ISBN 978-3-319-71323-6 ISBN 978-3-319-71324-3 (eBook)
https://doi.org/10.1007/978-3-319-71324-3

Library of Congress Control Number: 2018936135

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

It is my pleasure to introduce this volume "Methods and Biostatistics in Oncology—Understanding Clinical Research as an Applied Tool." It is a timely contribution to clinical research in oncology as we experience an unprecedented increase in the number of clinical scientists and clinical studies all around the world. Important advances in our understanding of several topics, such as genomics, immunology, targeted treatments, and biomarkers, capture headlines in the popular press. As translational efforts to use these advances in the clinic intensify, the need for appropriate clinical research methodology has never been greater. It is for this reason that this volume will serve an important need.

Dr. Araújo's and Dr. Riechelmann's editing skills are apparent in the team of authors they have recruited and the topics they have chosen. Each of the book's 20 chapters has been written by one or more internationally recognized experts in the field. The chapters cover substantial ground, starting from a historical introduction and moving on to study design. Several important technical aspects, such as the interpretations of multivariate analysis and survival analysis, and descriptions of case-control and cohort studies, are aptly included. Clinical trial design receives the attention it deserves and emerging fields such as cost-effectiveness and patient-reported outcomes are also included. These chapters can be read sequentially like a textbook, which will be valuable for those who are in the early years of their clinical research training or careers. It would be a mistake, however, to think that seasoned investigators will not benefit from this volume. Since most clinical researchers lack comprehensive formal training in methodology, their knowledge of statistical methods is limited. I would urge them to pick up this volume and read the chapters that interest them. They will find that the chapters are self-contained and not demanding.

My primary advice to the reader is to come back to the chapters as they need to apply the material to their own work. This may, most commonly, be the research they are engaged in. If preclinical work has yielded a result ready for clinical testing, rereading the chapter "How to Design Phase I Trials in Oncology" will reveal the subtleties of the presentation and will, no doubt, increase retention of the knowledge. There are more opportunities to relate this material to work, however. If one is refereeing a paper and the survival analyses are puzzling, if one is mentoring a student who is struggling with a multivariate analysis, or if one is reading an article with a cost-effectiveness analysis there will be much to learn by visiting the relevant chapters.

This book also stands apart because it has a separate chapter for bias (my favorite topic). The word originates from Bias of Priene, one of the seven sages of ancient Greece who thought and wrote a great deal about justice and fairness. It is ironic that we use his name to refer to certain types of prejudice and, in the scientific context, a systematic dissonance between the findings and the truth. Bias is widely recognized as a threat to the validity of a study, to the point that several types of commonly encountered biases have earned their own names, such as selection bias, verification bias, recall bias, etc. Bias is possibly the single most important concept in research methods and yet it might be the most misunderstood one. Bias usually arises from systematic differences between the sample analyzed and the population for which conclusions are drawn. Bias can be due to deficiencies in design; inadequacies in data collection; and legal, ethical, or other constraints. I am heartened to see bias receiving coverage in this volume, because it is even more important to consider bias in the age of big data, where automated data collection and the ability to merge disparate data sources leads to a huge amount of observational data and also makes it more difficult to understand what sorts of biases might have crept in. Some well-publicized failures such as Google Flu Trends and the Boston Pothole Experiment point to the importance and difficulty of detecting biases.

Let me finally make the point that the understanding of research methodology and statistics remains challenging, requiring great intellect and creativity. There is a big gap between results obtained by pushing a button or executing a command in data analysis software, and gaining knowledge and insight from these results, and this gap can be closed only by having a good grasp of research methodology. That is why you should read this book and recommend it to others.

New York, NY, USA Mithat Gönen

Contents

Contributors

Pedro Aguiar Jr., M.D. Faculdade de Medicina do ABC, Santo André, SP, Brazil

Roberto Jun Arai, M.Sc., Ph.D. Clinical Research Unit, State of São Paulo Cancer Institute, Faculty of Medicine, University of São Paulo, São Paulo, SP, Brazil

Raphael L. C. Araújo, M.D., Ph.D. Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos Cancer Hospital, Barretos, SP, Brazil

Carmelia Maria Noia Barreto, M.D. Clinical Oncology Sector, Sociedade Beneficência Portuguesa de São Paulo, São Paulo, SP, Brazil

David Bristol, Ph.D. Independent Consultant, Wiston-Salem, NC, USA

Brittany L. Bychkovsky, M.D., M.Sc. Dana-Farber Cancer Institute, Boston, MA, USA

Harvard Medical School, Boston, MA, USA

Samantha-Jo Caetano, M.Sc. McMaster University, Hamilton, ON, Canada

Louise Carter, M.A., M.B.B.S., Ph.D., M.R.C.P. The Christie NHS Foundation Trust, Manchester, UK

Division of Cancer Sciences, Faculty of Biology, Medicine and Health, Manchester, University of Manchester, Manchester, UK

André Lopes Carvalho, M.D., Ph.D., M.P.H. Teaching and Research Institute, Barretos Cancer Hospital, Barretos, SP, Brazil

Gilberto de Castro Jr., M.D., Ph.D. ICESP—Medicine School of University of São Paulo, São Paulo, SP, Brazil

Sirio Libanês Hospital, São Paulo, SP, Brazil

Natalie Cook, M.B.Ch.B., M.R.C.P., Ph.D. The Christie NHS Foundation Trust, Manchester, UK

Division of Cancer Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Beatriz Teixeira Costa Laboratory of Neuromodulation, Center of Clinical Research Learning, Spaulding Rehabilitation Hospital, Harvard Medical School, Boston, MA, USA

Emma Dean, B.MedSci.,B.M.B.S.,PhD.,F.R.C.P. The Christie NHS Foundation Trust, Manchester, UK

Division of Cancer Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Early Clinical Development, Oncology Translational Medicine Unit, Astra Zeneca, Melbourn, Hertfordshire, UK

Diane L. Fairclough, Dr.P.H. Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, USA

Isadora Santos Ferreira Laboratory of Neuromodulation, Center of Clinical Research Learning, Spaulding Rehabilitation Hospital, Harvard Medical School, Boston, MA, USA

Felipe Fregni, M.D., Ph.D., M.P.H. Department of Physical Medicine and Rehabilitation, Laboratory of Neuromodulation, Center of Clinical Research Learning, Spaulding Rehabilitation Hospital, Harvard Medical School, Boston, MA, USA

Mina Georgieva, M.S. Georgia Institute of Technology, Atlanta, GA, USA

Debra A. Goldman, M.S. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Rodrigo Santa C. Guindalini, M.D., Ph.D. CLION, CAM Group, Salvador, BA, Brazil

Department of Radiology and Oncology, State of São Paulo Cancer Institute, Faculty of Medicine of the University of São Paulo, São Paulo, SP, Brazil

Benjamin Haaland, Ph.D. University of Utah, Salt Lake City, UT, USA

Axel Hinke, Ph.D. CCRC, Düsseldorf, Germany

WiSP Research Institute, Langenfeld, Germany

Roxanne E. Jensen, Ph.D. Department of Oncology, Georgetown University, Washington, DC, USA

Bellinda L. King-Kallimanis, M.Sc., Ph.D. Pharmerit International, Boston, MA, USA

T. Peter Kingham, M.D. Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Monika K. Krzyzanowska, M.D., M.P.H., F.R.C.P.C. Division of Medical Oncology & Hematology, Princess Margaret Cancer Centre, Toronto, ON, Canada

Gilberto de Lima Lopes Jr., M.D., M.B.A., F.A.M.S. Global Oncology Program, Sylvester Comprehensive Cancer Center at the University of Miami, Miami, FL, USA

Ciara O'Brien, B.Sc., M.B.B.S., Ph.D. The Christie NHS Hospitals Trust, Manchester, UK

Cleyton Zanardo de Oliveira, M.Sc. Teaching and Research Institute, Barretos Cancer Hospital, Barretos, SP, Brazil

Education and Research, BP - A Beneficência Portuguesa de São Paulo, São Paulo, SP, Brazil

Katherine S. Panageas, Dr.P.H. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Bruno S. Paolino, M.D., Ph.D. Department of Cardiology, State University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Allan A. Lima Pereira, M.D., Ph.D. Department of Gastrointestinal Medical Oncology, University of Texas - M.D. Anderson Cancer Center, Houston, TX, USA

Laura C. Pinheiro, M.P.H., Ph.D. Division of General Internal Medicine, Weill Cornell Medicine, New York, NY, USA

Gregory R. Pond, Ph.D., P.Stat. McMaster University, Hamilton, ON, Canada

Melanie Powis, M.Sc. Division of Medical Oncology & Hematology, Princess Margaret Cancer Centre, Toronto, ON, Canada

Rachel P. Riechelmann, M.D., Ph.D. Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil

Everardo D. Saad, M.D. Dendrix Research, São Paulo, SP, Brazil

IDDI, Louvain-la-Neuve, Belgium

Andre Deeke Sasse, M.D., Ph.D. Department of Internal Medicine, Faculty of Medical Sciences, University of Campinas (UNICAMP), Campinas, SP, Brazil

Deise Uema, M.D. ICESP—Medicine School of University of São Paulo, São Paulo, SP, Brazil

Sirio Libanês Hospital, São Paulo, SP, Brazil

Fabiana de Lima Vazquez, D.D.S., Ph.D. Teaching and Research Institute, Barretos Cancer Hospital, Barretos, SP, Brazil

Vinicius de Lima Vazquez, M.D., Ph.D. Department of Surgery, Sarcoma and Melanoma Unity, Teaching and Research Institute/IEP and Molecular Oncology Research Center/CPOM, Barretos Cancer Hospital, Barretos, SP, Brazil

Francisco Emilio Vera-Badillo, M.D., M.Sc. Canadian Cancer Trials Group, Queen's University, Kingston, ON, Canada

Centro Universitario Contra el Cancer, Hospital Universitario Dr. Jose E. Gonzalez, Universidad Autonoma de Nuevo Leon, Monterrey, Mexico

Cheng Tzu Yen, M.D. Oswaldo Cruz German Hospital, São Paulo, SP, Brazil

Brief History of the Scientific Method and Its Application in Oncology

1

Vinicius de Lima Vazquez and Raphael L. C. Araújo

1.1 Ancient Science

The first step taken on the way to the scientific method was the cognitive revolution that occurred in our species nearly 70,000 years ago. Within the domain of language, it was possible to create vast collaborations among individuals, with abstract common values. This cognition allowed us to make the first attempts to explain our world, with myths and gods, many of them anthropomorphic. After the agricultural revolution, circa 10,000 BCE, and more recently (1000–500 BCE) with the development of writing and the rise of political and monetary systems, the first attempts to explain nature in a systematic way, and not in supernatural terms, began to flourish.

The first Western thinkers arose in ancient Greece and they utilized the observation of natural phenomena in developing theories based on those observations. Thales of Miletus, one of these pioneers, theorized that water was the origin of all forms in the universe.

The evolution of ancient Greek science took about 700 years and was an astonishing example of how an organized society where free thinking and education are greatly valued can flourish, albeit that education was available only to a minority. This environment gave to humankind great philosophers who enormously influenced our past and present knowledge. Plato and Aristotle were the transcendent figures of their times. Plato gave us, among other brilliant concepts, the concept of dualism and the value of ideas or an ideal world, as well as theories of mathematics. Aristotle, on the other hand, valued observations and gave us rules of nature

V. de Lima Vazquez, M.D., Ph.D. (✉)
Department of Surgery, Sarcoma and Melanoma Unity, Teaching and Research Institute/IEP
and Molecular Oncology Research Center/CPOM, Barretos Cancer Hospital,
Barretos, SP, Brazil

R. L. C. Araújo, M.D., Ph.D.
Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos
Cancer Hospital, Barretos, SP, Brazil

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_1

summarized in a methodical way. His six-book collection on logic, *Organon*, which set the basis of rational enquiry, was a tool used for thinking about and understanding nature for more than one thousand years. Aristotle also proposed four kinds of causation in nature: matter (material cause), form (formal cause), agent (efficient cause), and end (final cause). The ancient philosophers also developed a four-element explanation of the constituents of nature (earth, water, air, and fire).

Medicine was a highly intellectual profession in ancient Greece and in the Roman Empire. Galen (Claudius Galenus 129–216 CE) was a prominent Roman physician and philosopher in his time. He made extensive anatomical observations and promoted the theory and the typology of human temperaments according to an imbalance in the four bodily fluids. In parallel to the four elements of nature, the four bodily fluids were regarded as black bile or melancholia, yellow bile, blood, and phlegm. According to his theory, diverse diseases with diverse features and severity would occur in relation to different imbalances. Galen's writings were followed for centuries and were apposite with the classical Greek fundamental idea that the universe is perfect and that what goes wrong is related to deviations of the universal proposal for all things.

1.2 The Middle Ages and the Arabic Influence

Christianity and the disruption of the Roman Empire transferred the development of scientific thought to the Arabic realm, which blossomed in the Middle East during the Middle Ages. Many sciences flourished in this region during that time, with the most dramatic discoveries arising from mathematics and medicine. Some of the lost ancient knowledge was secretly preserved in Catholic monasteries in Europe, while some ancient books were translated into Arabic, and were studied and interpreted in Arab lands as a basis of new discoveries. In the late Middle Ages, many such books traveled back to the West.

Aristotelian thought matched Christian theology. William of Ockham, Siger of Brabant, Boethius of Dacia, and, mainly, Thomas Aquinas (1225–1274) brought a rational approach toward understanding nature, which they saw as a divine creation. In Europe, universities such as the University of Oxford, the University of Paris, and the University of Padova, among the oldest of the European universities, were established and Aristotelian/Galenic thought became solid and traditional. The main assumption was that perfect wisdom belonged to the past, and it was understood that no effort was needed to generate knowledge, but the task was to learn and recover knowledge from the ancients.

1.3 The Renaissance and the New Scientific Method

The new world discoveries in the late fifteenth and early sixteenth centuries shook and irreversibly changed the European way of thinking. The heliocentrism of Copernicus, Galileo's telescope, and other theories and technologies showed a different version of the natural world. Scholastic Aristotelian thought was not relevant anymore.

In 1543, Andreas Vesalius, in his masterpiece *De Humani Corporis Fabrica*, demonstrated human anatomy in bright new colors. His work was the result of systematic and meticulous dissections of human corpses and showed many differences from the traditional anatomy of Galen. Different from Vesalius, Galen, obeying the Roman law, dissected monkeys, dogs, and other animals, but not humans. Further observations from William Harvey correctly described the circulation of the blood in humans, and the ancient fluid imbalance theory of Galen was disproved. A new way to explain and to explore the complexity of the world was necessary. Could a suitable method be found?

1.3.1 The "Magic" World and Natural Philosophy

Since the Middle Ages, there had been a "magical" way of observing and classifying knowledge related to practical and unexplained phenomena, using methods such as alchemy and theories of magnetism, among others. During the Renaissance, for the first time, these natural or manipulated phenomena started to attract intellectual attention and attempts were made to explain what were previously considered as curiosities or bizarre happenings with occult and supernatural causes by reference to the same forces or laws conceived as governing all of nature. One beautiful example of such intellectual examination was the treatise *De Magnete* (1600), by William Gilbert, where a very well-known technology, utilized by sailors of the time for navigation, was depicted in detail and where Gilbert concluded, in very demonstrative and elegant form, that the Earth is similar to an enormous magnet.

Francis Bacon, in a world very confused by the rupture of the scholastic model, proposed a new method of natural philosophy, later called natural science, with branches such as biology and all life sciences. His *Novum Organum* (new instrument) (1620) was ambitiously intended as a substitute for Aristotle's *Organon*. Briefly, Bacon defined empirical methods to explain nature, using induction after real observations (empiricism) instead of deduction (which is supported by impalpable and weak elements). This became known as the Baconian method; with this method, observations must be as extensive as possible to rule out unexpected manifestations, and the simplest explanation of causation should be sought. Interestingly, the method opened possibilities of new answers for old questions.

Rene Descartes was another great thinker who modeled our methods in sciences. In French society, where skepticism was growing as an answer to the absence of reliable ways to understand nature, he framed his thoughts and arguments to resist the skeptics' attacks. His famous statement *Cogito ergo sum*, translated as "I think, therefore I am", is much better interpreted as "I doubt, or I question, therefore I am". The doubt or question was the first and core principle of the four principles in his method, explained in his famous book *Discourse on the Method* (1637). The first principle can be explained as that a supposition would last only if it could stand after all questions have been asked. In his own words: "The first was never to accept anything for true which I did not clearly know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in my judgment than what was presented to my mind so clearly and distinctly as to exclude all ground of doubt". His other principles are still important in the methods of modern sciences. The second principle is "to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution"; the third, "to conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex; assigning in thought a certain order even to those objects which in their own nature do not stand in a relation of antecedence and sequence". And the last principle is "in every case to make enumerations so complete, and reviews so general, that I might be assured that nothing was omitted." Descartes was also a brilliant mathematician, and mathematics is part of the understanding of his method. For him, the sharpness of calculus should be applied to the methods of science, for the precision of results and the search for truth. He contributed to the use of a methodology as an important point to both prove and reproduce experiments. More than this, according to him, precision was the only way to achieve answers. These concepts were powerful and still resound nowadays.

Sir Isaac Newton (1643–1727) was one of the most prominent scientists in human history. His discoveries, in mechanics, optics, mathematics, and other fields, were revolutionary. He demonstrated, with the power of mathematics, the classic laws of mechanics and this was well aligned with the methods of Descartes, paving the way for the modern scientific method.

1.4 The Industrial Revolution and the Birth of Clinical Cancer Research

Modern medicine and the rise of contemporary oncology and clinical cancer research were shaped, as we know today, after the industrial revolution of the nineteenth century. The technology acquired during this period allowed new discoveries to be made and opened possibilities for surgery, radiation therapy, and more recently, the use of antineoplastic drugs.

James Lind from Scotland is considered to be the first physician to have conducted a clinical trial. On a ship, in 1747, when trying to treat widespread morbid scurvy, he designed a comparative study in which twelve sailors with scurvy were allocated to two groups, each with a different diet complement every day. When Lind observed the results, he found that the consumption of oranges and lemons (sources of vitamin C) led to cure of the disease.

The idea of the placebo arrived in the 1800s. Dr. Austin Flint, in the United States, when studying rheumatism, had the idea of giving a herbal "placebo" compound instead of an established medicine. He published details of this experiment in his book *A Treatise on the Principles and Practice of Medicine* (1866). Some patients receiving the "placebo" compound actually improved and he concluded that this was because of the confidence patients had in the treatment they believed they were receiving.

Controlled, blinded, and '*a posteriori*' randomized trials were first designed and conducted in the late 1940s and 1950s. During that time there were great advances in epidemiology and biostatistics. Ronald Ross, Janet Lane-Claypon, Anderson Gray McKendrick, and others introduced the mathematical method in epidemiology. In the field of oncology, the seminal work of Doll and Hill, in the British Doctors Study (1956), introduced the statistical concept of the hazard ratio and proved that tobacco consumption led to a higher risk of lung cancer. All these concepts and tools in research methods were crucial to the development of modern oncology.

Further, the two world wars had a great impact on the rise of clinical research. The unethical human experiments performed before and during World War II endorsed by the German government were the result of a policy of racial hygiene, in pursuit of a pure "Aryan master race". In 1947, during the Nuremberg War Crimes trials, the Nuremberg Code was established. The Code states that, for any research in humans, the subjects of the research must give their full consent and participate voluntarily, and there must be no unnecessary or unsafe exposure of the participants to any agent or procedure. The Code was based on the principle of giving benefit and doing no harm to the participants. Almost 20 years later, in 1964, the Declaration of Helsinki was made by the World Medical Association and this provided another cornerstone in clinical research practice. The Declaration holds that all research in humans should be based on a scientific background, with putatively more benefits than risks; new treatments should be compared with actual standard treatments, and approval of the project must be obtained from an independent committee of ethics in research (for instance, an institutional review board); there must also be a declaration of any conflict of interest, among other factors.

Historically, the first curative treatment for cancer was surgery. Although it started empirically, far from the modern methods, based on hits and misses, it represented a fantastic advance for modern medicine. At the end of the nineteenth century, William Stewart Halsted, acclaimed by many as the father of surgical oncology, systematically observed the results of his mastectomy surgeries. He noticed that recurrences occurred in a very predictable pattern and he created a new surgical technique, which included a more aggressive approach with resection of the pectoral muscles and the lymphatic nodes from the axilla. This radical (from the Latin word meaning "root") surgery became a model for oncological surgery overall for over a century, and "en-bloc" or "radical" resection remains as a common concept in surgical oncology. Only after the evolution of clinical research methods in the 1980s were new less aggressive surgical methods proposed, showing lower morbidity. These new surgical methods were accepted only because comparative studies demonstrated they were superior to the Halsted methods, with undeniable proof of benefit. For example, Umberto Veronesi, in Italy, and Bernard Fisher, in the United States, conducted randomized trials where they demonstrated that, for localized small breast cancer, local tumor control could be achieved with less aggressive surgery, adjuvant chemotherapy, and radiation therapy instead of total mastectomy.

This led to a paradigm shift in the idea of cancer treatment being exclusively surgical. Fisher proposed that when breast cancer presented an early hematogenous spread, then the lymph node involvement would simply represent systemic disease and not only locally advanced disease. This was the rationale for associated adjuvant treatment after breast surgery. Veronesi advocated breast-conserving surgery associated with adjuvant radiotherapy for local control, as well as chemotherapy for systemic treatment. Both these surgeons emphasized the importance of a multidisciplinary team for an oncological approach in treating cancer patients.

The use of radiation in medicine had a different course; it was described at the end of the nineteenth century by Pierre and Marie Curie, and it was used in oncology in the late 1930s to treat head and neck cancers. In the 1950s the use of cobalt teletherapy offered local treatment for many kinds of cancers. Radiotherapy also progressed to conserving techniques as the technology evolved to deliver doses with more precise techniques made available to give radiation to the target with less toxicity in the path of energy into the tissues.

The use of prospective controlled randomized trials is a milestone for clinical cancer research and for determining the standard treatment. The first such trial was conducted in breast cancer patients in 1968, comparing radical mastectomy (Halsted procedure) associated with thiotepa or placebo. This and other studies in breast cancer were carried out by a multicenter cooperative now called the National Surgical Adjuvant Breast and Bowel Project (NSABP), which Fisher led. These studies showed that better oncologic outcomes could be achieved by using a less radical procedure associated with adjuvant chemo- and radiotherapy, with less morbidity shown as well. With advances in the identification and stratification of clinical presentations of tumors, including clinical and demographic variations among individuals, multicenter trials became increasingly important. Increases in the sizes of study populations and the design and application of large phase III and IV studies clarified the effects of interventions over larger populations and showed more safety. To speed up the long and meticulous process of patient accrual, some studies became international, with dozens of centers involved.

Another successful advance in clarifying the methods involved in research in medicine and oncology was the concept of evidence-based medicine, developed in the 1990s. The scientific evidence of a new treatment or method could now be classified hierarchically according to different evidence levels (Fig. 1.1). This idea spread widely and became an important instrument for clarifying and showing the explicit quality of the research methods utilized in each study.

In parallel with this unprecedented advance, concerns about safety and ethics began to grow. To resolve such diverse concerns, other methods were introduced in the study designs to increase the safety of the participants. Safety monitoring and ethics committees were established, and good clinical practice guidelines policies, as well as a statistical calculus for endpoints and futility or early results presentation, became obligatory parts of experimental research in human beings. More recently, different initiatives have provided guidelines for methods of conducting clinical trials. Scientific and medical journals, government agencies and grants supporters require researchers to follow these guidelines. Some examples are shown in Table 1.1.

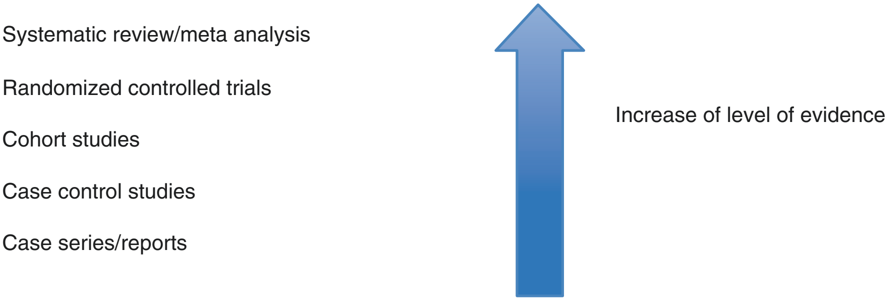

Systematic review/meta analysis
Randomized controlled trials
Cohort studies
Case control studies
Case series/reports
Increase of level of evidence

Fig. 1.1 Study design according to evidence-based relevance

Table 1.1 Methodological guidelines indicated for medical research, according to study design

Study design	Guideline
Clinical trial	CONSORT statement and EQUATOR http://www.consort-statement.org http://www.equator-network.org
Epidemiology, qualitative research, and mixed methods	STROBE and SRQR http://strobe-statement.org
Multivariable prediction models	TRIPOD http://www.tripod-statement.org
Routinely collected health data	RECORD http://www.record-statement.org
Systematic review	PRISMA http://www.prisma-statement.org
Quantitative PCR data	MIQE http://www.clinchem.org/content/55/4/611.long
Biomarker and association studies	REMARK http://www.nature.com/bjc/journal/v93/n4/full/6602678a.html

PCR polymerase chain reaction

1.5 The Future

The future of clinical research in oncology is a fascinating matter. Many recent advances in molecular methods and the immune landscape of tumors are bringing complexity to a whole new level. The implementation of next-generation sequencing and the massive output of genome, transcriptome, proteomic, and other molecular data, added to demographic and clinical data—exchanged and collected in the form of multi-institutional data for hundreds or thousands of individuals, some freely available in public consortia—are a challenge for the understanding of big data. Massive amounts of personal information have been collected and are available in real time, and this data, combined with results from the treatment of thousands of patients outside of clinical trials (where the treatment became approved), has led to new visions and new post-approval evaluation of treatments in "real patients", since the patients usually included in clinical trials have to be of good general clinical status, and clinical situations away from the mean are, in most cases, excluded, for bias control. This colossal data is merging into a new fascinating frontier in oncology: molecular targeted therapy, which, added to new immunology discoveries, provides more personal and precise treatments for patients. New tools in bioinformatics have emerged to enable a search for new solutions to speed up and improve the accuracy of diagnostic and therapeutic interventions.

However, the present methods in use cannot answer many questions prompted by the myriad information gathered. We still do not know the answers we dreamed of in relation to human gene and molecular discoveries for shaping the promise of personalized medicine for each individual cancer patient. Nevertheless, the recent astonishing advances in communication, computation, and artificial intelligence suggest that we are certainly living in a fantastic new era where new—hitherto inconceivable—discoveries may be realized within a lifetime. New approaches in methodology are warranted, and for certain these will arise in the near future.

Further Reading

1. Bhatt A. Evolution of clinical research: a history before and beyond James Lind. Perspect Clin Res. 2010;1(1):6–10.
2. Descartes R, Sutcliffe F. Discourse on method and the meditations. London: Penguin; 1968.
3. DeVita VT Jr, Rosenberg SA. Two hundred years of cancer research. N Engl J Med. 2012;366(23):2207–14.
4. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking. Br Med J. 1956;2(5001):1071.
5. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br Med J. 2008;336(7650):924.
6. Harari YN. Sapiens: a brief history of humankind. Toronto, ON: Random House; 2014.
7. Henry J. The scientific revolution and the origins of modern science. New York, NY: Palgrave Macmillan; 2008.
8. Henry J. A short history of scientific thought. Basingstoke: Palgrave Macmillan; 2011.
9. Lindberg DC. The beginnings of Western science: the European scientific tradition in philosophical, religious, and institutional context, prehistory to AD 1450. Chicago, IL: University of Chicago Press; 2010.

2 Generating a Hypothesis for an Oncology Study

Beatriz Teixeira Costa, Isadora Santos Ferreira,
and Felipe Fregni

2.1 Introduction

Clinical practice has long been recognized as a profession that combines clinical expertise with scientific evidence. At present, the need to be constantly updated while looking for new alternatives to improve patients' outcomes has transformed clinical research into an essential instrument for healthcare providers. However, the applicability of research findings to routine clinical practice remains incredibly challenging, as it requires in-depth knowledge and critical thinking.

Although the increasing number of studies throughout the past few decades has resulted in a positive impact on patients' lives, many questions remain unanswered. The eager need for breakthroughs and new ideas often makes researchers believe in a great number of misleading studies that do not take into account the basic aspects that characterize well-conducted research. Therefore, the relevance and validity of the studies must be a primary concern, considering that a great amount of scientific data does not always reflect high-quality information.

One of the first steps to be taken in conducting valid clinical research is to ask answerable and interesting questions. At a primary stage, it is fundamental to select a broad topic of interest and deeply explore the available literature in order to draw a line between the existing knowledge and the unknown. A review of published studies allows the recognition of current missing information, also

B. T. Costa • I. S. Ferreira
Laboratory of Neuromodulation, Center of Clinical Research Learning, Spaulding Rehabilitation Hospital, Harvard Medical School, Boston, MA, USA
e-mail: bcosta@neuromodulationlab.org; iferreira@neuromodulationlab.org

F. Fregni, M.D., Ph.D., M.P.H. (✉)
Department of Physical Medicine and Rehabilitation, Laboratory of Neuromodulation, Center of Clinical Research Learning, Spaulding Rehabilitation Hospital, Harvard Medical School, Boston, MA, USA
e-mail: fregni.felipe@mgh.harvard.edu

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_2

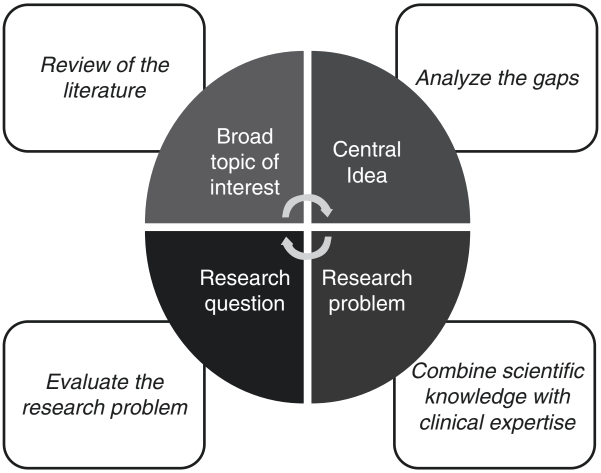


Fig. 2.1 Representation of the process for achieving a good research question

called gaps, and provides an essential rationale for identifying specific issues of importance. Also, it is useful to check clinical trials registrations to search and see what is being researched in the field. Throughout this process, the focus starts to convert a broad topic to a central idea. Thereafter, the combination of scientific knowledge with clinical expertise can be translated into a significant research problem (Fig. 2.1).

Choosing a relevant research problem in oncology is imperative for successfully conducting a study related to preventive, diagnostic, or therapeutic approaches. The challenging task, in fact, is to formulate a precise research question that contributes to and is complementary to the science in oncology. Also, the difficulty lies in finding questions that are simultaneously feasible and interesting. Indeed, questions can also be classified as low- and high-risk questions. Nonetheless, recent advances in oncology research have become a powerful incentive to overcome these barriers, thus leading to the development of well-conducted new studies and making progress against cancer.

2.2 Defining a Research Question

The process of defining a researchable question begins with the evaluation of a research problem and is directly related to the researcher's familiarity with a certain topic. Ideally, the question should be clearly stated at the end of the Introduction and must be as specific as the knowledge the investigator wants to gain. This will allow the investigator to properly answer the question within a given time interval. In addition, it is important to keep in mind that a research question must reflect what the investigator wants to know, an uncertainty about a problem that can be analyzed to obtain useful information.

Although the process is time-consuming and resource-demanding, the researcher should be passionate about the investigation and believe that it is worthwhile in order to fuel the work. In other words, an investigator must believe that, by answering a new research question, useful information will be generated and advances will emerge as a result, regardless of whether the results are in favor of or against the null hypothesis. A well-designed research question should be able to pass the “so what?” test, which indicates how meaningful a research question actually is.

After addressing the importance of a research question, it is fundamental to make sure that it is relevant to both the scientific community and the public, while also meeting certain criteria: it must be answerable and feasible, while increasing knowledge in the field. Even though several aspects must be taken into consideration to achieve an adequate research question, it is indispensable to determine the clinical concerns that should be explored while rationalizing the need for the investigation.

A useful tool suggested by Hulley and colleagues [1] that may guide the development of a successful research question is the FINER criteria (Table 2.1), the use of which increases the potential of developing a publishable study by summarizing five necessary main topics that should be outlined. Accordingly, a research question must be *Feasible*, *Interesting*, *Novel*, *Ethical*, and *Relevant*. In regard to *Feasibility*, the question must address an adequate number of subjects and the researcher must have adequate expertise; the study must be affordable in both time and money, and be manageable in scope. It should be *Interesting* enough to intrigue the investigator, peers, and the community, as well as being a *Novel* source of information that confirms, refutes, or extends previous findings. In addition, it must be *Ethical* so as to preserve the patients’ welfare, consequently receiving the institutional review

Table 2.1 FINER criteria for a good research question

F	Feasible	• Include an adequate number of subjects • Follow adequate technical expertise • Be affordable in time and Money • Manageable in scope
I	Interesting	• The answer should intrigue the investigator, peers and scienticic community
N	*Novel*	• Must confirm, refute or extend previous findings
E	*Ethical*	• Amenable to a study that institutional review board will approve
R	*Relevant*	• To scientific knowledge • Clinical and health policy • To future research

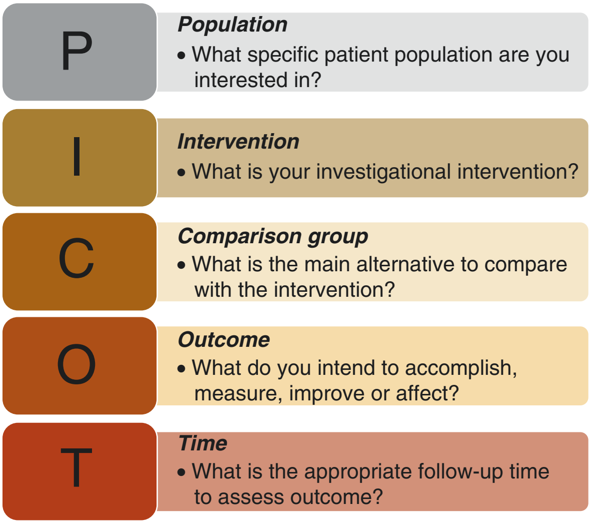

P
Population
• What specific patient population are you interested in?
I
Intervention
• What is your investigational intervention?
C
Comparison group
• What is the main alternative to compare with the intervention?
O
Outcome
• What do you intend to accomplish, measure, improve or affect?
T
Time
• What is the appropriate follow-up time to assess outcome?

Fig. 2.2 PICOT format for developing a research question

board's (IRB's) approval, and *Relevant* to scientific knowledge, clinical and health policy, and finally to future research.

Whereas the FINER criteria address general aspects of the research question, the PICO format, often mentioned in the literature as the PICOT format (Fig. 2.2), increases the investigator's awareness of the important aspects to mention in the research question, such as the specific *Population* of interest (main criterion). Moreover, this helpful format outlines the effects of a certain *Intervention* by describing the *Comparison group*, *Outcome of interest*, and the amount of *Time* required to assess the outcome:

2.2.1 P (Population)

Population represents the sample of subjects to be recruited for a study; individuals in whom the knowledge is required. For instance, in a study in which the purpose was to "identify circulating microRNAs able to identify ovarian cancer patients at high risk for relapse [2]", the population was *cancer patients in high risk for relapse*.

It is essential to remember that it is not often easy to determine a sample that is most likely to respond to an intervention (e.g., absence of metastasis) and one that can be generalized to patients that are more likely to be identified in daily practice. Other considerations can be dictated by the availability of patients. By addressing questions such as: What is the appropriate age range? Should males and females be included? What about co-morbidities? And Is the type of tumor a relevant factor?, the researcher is able to narrow down the group of individuals that would be the main focus of the study.

One of the key factors to be aware of before defining the population of interest, along with the inclusion and exclusion criteria, is the potential risk of bias, the internal validity of the results as well as their generalizability. The more rigorous the inclusion and exclusion criteria, and thus the more restricted the target population, the greater their influence on the applicability of the results. Although a restricted population may reduce the risk of null results, thus increasing internal validity, it might also considerably diminish the generalizability of the study. On the other hand, despite representing patients seen in daily practice, a broader population and broader inclusion criteria may have the opposite effect.

Hence, an inadequate definition of the criteria that will shape the population of interest may , as a result, alter the study design, leading to unsuccessful findings and decreasing the chances of achieving clinical significance.

2.2.2 I (Intervention)

Also referred to as exposure, “I” corresponds to the treatment, procedure, therapy, or placebo that will be provided to the patients enrolled in the study. In a study that aims to “evaluate the security and effectiveness of cisplatin with constant dose-intense temozolomide (TMZ) for reduplicative glioblastoma multiforme (GBM) within 6 months” [3], for example, the intervention would be *cisplatin plus constant dose-intense TMZ.* Before designating an intervention to a group of patients, it is important to take into account previous studies, if existent, so as to predict estimates of the study’s effect. The following step after ensuring the safety of the exposure is to define, in advance, how to measure its efficacy: using clinical outcomes, surrogates (biomarkers), questionnaires, quality-of-life scales, or other methods. Finally, it is necessary to analyze the financial aspects involved in a certain intervention, counterbalancing its pros and cons, as well as analyzing its cost-effectiveness.

2.2.3 C (Comparison Group)

The comparison group is a group of subjects that resembles the experimental group in several aspects, but who do not receive the active treatment under study. Control interventions may be in the form of a placebo, standard care or practice, a different therapy, or even no intervention. A clear example of “C” is in a study by Middleton et al. [4] that aimed to investigate the clinical efficacy of “vandetanib plus gemcitabine versus placebo plus gemcitabine in locally advanced or metastatic pancreatic adenocarcinoma”. In this study, while the active group received vandetanib and gemcitabine, the comparison group received placebo and gemcitabine, which was the standard treatment at that time.

Additionally, a research question could likely change depending on the control groups. In other words, a question that aims to compare one intervention versus another is different from a question that aims to compare one intervention versus no intervention. Therefore, the comparison between groups has large implications in a study and is intrinsically associated with the process of defining participants’ exposure to the intervention.

2.2.4 O (Outcome of Interest)

The outcome corresponds to a variable or a result that can be measured in order to examine the effectiveness of an intervention. A number of outcomes can be chosen, with the most commonly used outcomes in oncology being the objective response rate, progression-free survival, overall survival, and patient-reported outcomes. The investigator needs to understand what is the (single) outcome that can best measure the effects of a given intervention. A study by Shipley et al. [5] exemplifies an outcome that can be used in research. This study aimed to evaluate whether antiandrogen therapy plus radiotherapy would further improve cancer control as a salvage therapy for recurrent localized prostate cancer and prolong overall survival in comparison with radiation therapy alone. The study used the overall survival rate as a primary outcome.

Importantly, the study objective is not the same as the study endpoint. An objective refers to *what* we want to find, while an endpoint defines *how* we will find the outcome. For example, a phase II trial aims to evaluate the efficacy of a new oral chemotherapy agent; in this case, the study objective is to determine drug efficacy, while the primary endpoint can be progression-free survival from the date of the first dose of treatment or the response rate at 8–12 weeks from treatment initiation, for instance.

As the selection of an outcome plays a major role in the interpretability of the research question, it is necessary to consider the variations in order to formulate a reliable and well-structured endpoint. Furthermore, the main outcome should reflect the measurement of the most significant aspects related to the patient's condition, being sensitive to the effects of the intervention at the same time. Finally, an ideal endpoint should be reachable so that the investigator is able to measure and assess it. For instance, if an invasive procedure is required to obtain a wanted outcome (e.g., histological aspect of a tumor), the feasibility of using the procedure needs to be considered as compared with other options.

The outcomes of a study can be mainly divided into primary and secondary. The primary outcome is the one of main importance; in other words, the one that will guide the main study design and will be used for sample-size calculation. The secondary endpoints are the ones used to investigate additional effects of the intervention. These should also be pre-established in the protocol, stating their relevance and why they are important for the trial. Predetermining secondary outcomes increases the strength of secondary findings.

Researchers often face the dilemma of deciding which type of outcome to pursue: a clinical or a surrogate (biomarker). There is still much controversy about this matter; however, regardless of the choice, these concepts are well established. A clinical outcome must be a clinical event that is closely related to the patient or a direct measurement of how a patient feels or behaves. On the other hand, a surrogate outcome is an indirect measurement, usually of biomarkers, which incorporates laboratory measures, radiological examinations, physical signs of a disease, and other factors. Although it is an indirect measure, in order to be valid, the biomarker should be able to predict a clinical outcome that represents a true benefit to patients.

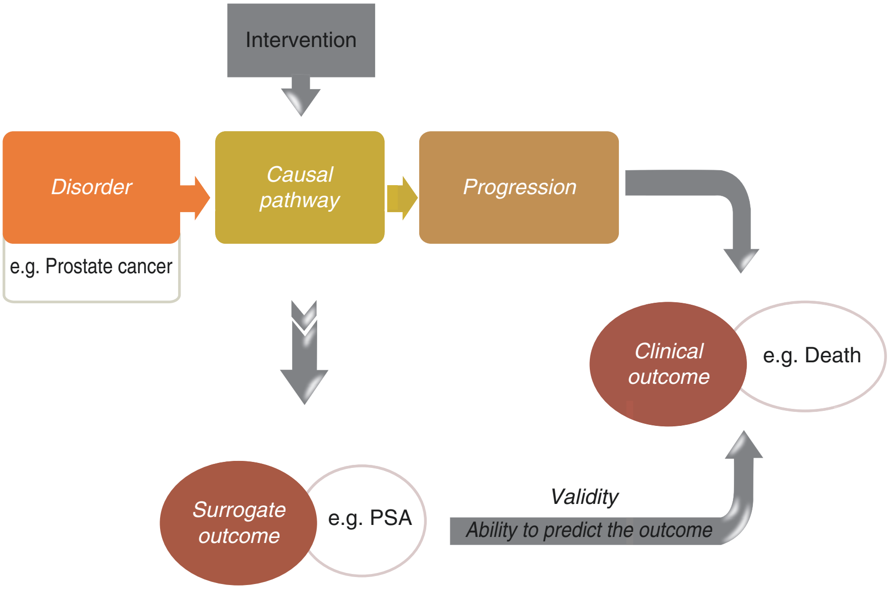


Fig. 2.3 The process for achieving surrogate and clinical outcomes

In oncology, a true endpoint is overall survival, while a surrogate endpoint can be time to tumor progression, for example.

In oncology, tumor markers such as prostate specific antigen (PSA) are other examples of surrogate outcomes. Some of these markers have been correlated with clinical efficacy measures, in this case death or symptoms, as they may provide information about the disease progression. Nonetheless, alterations in the biomarker may generate unreliable information, because the causal pathway, which directly indicates the risk of mortality, is that of the tumor itself (Fig. 2.3).

For this reason, it is important to consider the advantages and disadvantages of surrogate vs. clinical outcomes. For further discussion of the pros and cons of utilizing surrogate endpoints of overall survival in cancer trials, refer to Chaps. 3 and 7.

2.2.5 T (Time)

Time describes the duration of the data collection and the follow-up period for the main event; for example, tumor progression, to arise. For instance, in a study by Guan et al. [6], which intended to compare “intensity-modulated radiotherapy with or without weekly cisplatin for the treatment of locally recurrent nasopharyngeal carcinoma between April 2002 and January 2008”, the enrollment time was from April 2002 to January 2008, with a median follow-up time of 35 months per patient, ranging from 2 to 112 months.

In the literature, it is possible to find some references that do not include this parameter as part of the research question. However, it is important for the investigator to know, from the beginning of the study planning, that the period between data collection and the end of the expected follow-up duration will correspond to the main endpoint. In oncology studies, it is important that investigators acknowledge tumor biological behavior so as to adequately estimate the follow-up period. For example, for indolent metastatic tumors, such as well differentiated neuroendocrine tumors, it may be necessary to follow patients for months to years in order to observe deaths; in contrast, in trials of second-line therapies for metastatic pancreatic cancer, the follow-up period for assessing survival is measured in weeks.

In oncology, another aspect to consider before defining the length of data collection is the risk of missing data, owing to dropouts, as this can be relatively high in some oncology trials, particularly in the refractory metastatic setting and in cancer survivorship cohorts.

Assuming that the following research question has been defined: *Is drug ABC more efficient in alleviating distress than drug XYZ in patients with cervical cancer within 2 years?*, how can a researcher ensure that the question meets the PICOT format? By testing:

P → *Patients with cervical cancer*
I → *Drug ABC*
C → *Drug XYZ*
O → *Distress*
T → *2 years*

All questions must be defined in the planning stages of the study and all additional questions must never compromise the primary one. In fact, the primary question should be the researcher's ultimate focus, as its relevance is strictly correlated with the generation of the basis for the study's following steps: hypothesis and study objectives. And always keep in mind to pre-define the study endpoint. In this hypothetical question, the outcome "distress" must be carefully detailed and determined *before* the study starts. For example, one must set *a-priori* which scale will be used to classify patients as "distressed", which validated instruments and cut-offs will be used, what change in the scale will be considered clinically relevant, what will be the intervals and timing of the questionnaire, and how the questionnaire will be applied to the study participants.

2.3 Developing a Strong Research Question for a Grant Application

The United States National Institutes of Health (NIH) supports research projects and encourages investigators to consider a number of aspects when developing a study. As part of the initial research process, extreme importance must be given to the significance of the study, its feasibility, and its potential for innovation in medical practice. These components must be carefully discussed among investigators so as to guarantee that the study is likely to have a significant impact and promote substantial scientific progress. Each of the three essential components are discussed in the following sections.

2.3.1 Significance

A significant study allows the advance of scientific knowledge, technical capability, and clinical practice. Conceptually, significance refers to the study's importance in the field and its potential contributions to current knowledge. After addressing an important problem, a work that has significance can provide valuable information to improve medical practice and thus have an impact on patients' welfare. In order to verify how significant a study can be, it is advisable to ask a few questions in the first stages of the research planning.

- *"Why is the study being conducted?"*
- *"Can the research findings change clinical practice?"*
- *"Does the study improve any aspect of people's lives?"*
- *"What could be the overall impact of the study in the medical community?"*
- *"Will the study have major effects in a given population?"*

By answering these questions, it is possible to understand why the study should be performed and what will be its implications for the scientific community and the general public. Therefore, researchers must reflect on whether the proposed work has relevance and will expand knowledge.

Determining the significance of a work may seem an effortless task. However, it is often difficult to distinguish a groundbreaking work from just an attractive theory. Despite the challenges, being aware of the study's significance is fundamental to ensure the study's credibility and to qualify it for funding opportunities in subsequent phases of the research process.

2.3.2 Innovation

The innovation component intends to specify which novel approach or method a proposed study employs. There is no reason to conduct research if no novel practice is explored or new methodologies are not prospectively developed. The actual goal of doing clinical research must be to advance the frontiers of understanding and develop new paradigms in order to benefit medical practice.

Researchers must seek originality and innovative approaches with a view to promoting a relevant study that will directly influence clinical practice. As oncology research is mostly related to preventive, diagnostic, and therapeutic approaches, the innovative factor in a study in this field is commonly related to one of these aspects.

2.3.3 Feasibility

Throughout the process of exploring a research problem and defining an answerable question, it is crucial to assess the feasibility of a given project. Conceptually, feasible studies are those that are capable of being conducted and likely to deliver success. In fact, researchers may eventually develop studies with significance and innovative potential, although they are unfeasible.

In order to ensure the feasibility of a study, numerous factors must be accounted for. The necessary *time* to successfully complete the study, thus having relevant results, is one of these factors. For instance, if the time of the study is underestimated, there is a chance of either recruiting an insufficient number of participants or not giving a long enough follow-up for patients to experience the event, which may compromise the study power. Another important factor is the amount of *resources* required to successfully conduct research. Indeed, owing to a lack of resources, several steps of a study may be highly affected, such as recruitment, the collection of data, and data analysis.

In oncology studies, the scenario regarding feasibility is no different from that of studies in other fields. The fundamental factors that must be contemplated are the same as the ones mentioned above. However, in this field, some factors demand more careful attention. For instance, the high cost associated with cancer drugs and molecular analyses is a significant concern when conducting oncology research. At present, as various clinical trials frequently investigate the effects of oncologic drugs, the success of this type of study commonly depends on industry sponsorship and a large amount of resources.

Finally, when evaluating the feasibility of a study, researchers must also identify the potential risks and benefits associated with the intervention. Addressing any predictable harm or balancing such harm with the expected benefits is an important consideration to be made at this stage of the process.

2.4 Risks Involved in Clinical Research

Clinical research invariably presents a source of risk to the participants, as any intervention in humans may directly or indirectly affect their lives. The concept of risk corresponds to the probability of some future harm or unwanted event occurring. In fact, potential risks involved in research can have different classifications, as humans may be harmed in the psychological, economic, physical, and social spheres. In order to prevent harm from occurring and to protect research subjects, it is mandatory to assess every foreseeable risk in the first stages of the study planning, while still deciding on the research question. This is especially important when it comes to knowing and adhering to all the rules and regulations required for human subject research.

When submitting a protocol to an IRB or ethics committee, one of the main concerns should be the relation between safety and risk. In order to proceed, the study must consider their subjects' safety as a high priority and make sure that a specific intervention is not harming the patient.

The correct assessment of the potential harms of a given intervention can be accomplished by searching available evidence, seeking experts' impressions, or considering clinical experience. The assessed risks that cannot be prevented during the study must then be justified and counterbalanced with the prospect of benefits.

In oncology, as well as in other fields, risks should be assessed before the study by determining their nature, likelihood, and severity. For instance, suppose that a researcher is planning to run a clinical trial to evaluate the effects of a new drug in patients who have lung cancer. What are the potential risks involved in this study? In this example, important aspects to consider are the discomfort of drug administration and the possible side effects of the drug, such as nausea, fatigue, or appetite loss, or even more severe effects, such as febrile neutropenia, pneumonitis, and cardiovascular events. Additionally, in the psychological sphere, there may be a risk of embarrassment to the patients when they answer questions about smoking habits and illicit drug use, depression episodes, or transient anxiety. Accordingly, all possible injuries must be accounted for early in the study, so that future consequences can be minimized.

2.5 Formulating a Research Hypothesis

The following step of the research process covers the formulation of study hypotheses, which are created from a precise research question. It is fundamental that this step follows the definition of the main problem and the study question, as these components will be the basis for establishing simple and specific hypotheses. In contrast, if hypothesis generation is not performed before the study starts, every potential conjecture will be biased by the subjects' performance and study results.

The concept of a hypothesis refers to a clear statement of the expected research outcomes. Investigators also define a hypothesis as a prediction of an intervention's consequence or as a tentative guess intended to guide the consecutive phases of a work. Hypotheses are extremely useful for identifying research objectives; defining abstract concepts of the study; driving data collection; and for establishing the relationship between the research problem and expected solutions and providing the assumptions for sample-size computation.

Undoubtedly, a research problem cannot be addressed without predicting the results of a study, and this justifies the necessity of generating a satisfactory hypothesis. However, researchers must not forget that hypotheses need to be valuable even when negative. In order to ensure this aspect, careful attention must be given to the construction of the research hypothesis itself. Specific tips are given as follows:

- A hypothesis must be clear and precise, without leaving ambiguities about the study outcomes.
- It should be testable. The statement should include the variables, the population, and the time.
- It should be written in concise language and in simple terms, in order to avoid misunderstandings.

- It must be neither too general nor too specific.
- It is crucial to write hypotheses in a declarative sentence form.

Example:

Patients with primary glioblastoma that undergo surgery followed by adjuvant radiotherapy and chemotherapy have longer survival than patients receiving only adjuvant radiotherapy.

Comment: In this example of a hypothesis, the study population and intervention are well defined, as are the expected results. Additionally, the hypothesis is written in a succinct declarative form, which gives a fair idea of the study goals.

In order to achieve an adequate research hypothesis, two possible methods can be used. The first one is to deductively define the research hypothesis, meaning that the *observation* would generate a *pattern*. Then, according to the pattern, a *temporary hypothesis* would be generated in order to define a *theory*. The second method is to inductively achieve the research hypothesis. In other words, a main *theory* would initially originate a *hypothesis*. After there has been an *observation*, the hypothesis might be confirmed or refuted. Thus, there is no right or wrong in this matter, but it is up to the researcher to choose which method is the most suitable.

2.6 Types of Hypotheses

2.6.1 Null Hypothesis

The null hypothesis, usually represented as H_0, represents the current state of knowledge. For instance, it states that there is no difference between groups in a statistical test. In fact, a study is conducted to find evidence to reject the null hypothesis. The researcher's main intention throughout the study is to reject the null hypothesis, meaning that the original idea that generated the investigation is correct. However, if a study fails to reject the null hypothesis, it does not mean that the study is doomed to failure. The researcher may opt to adjust the original rationale and develop a new study with a different research question so as to possibly reject the null hypothesis.

To exemplify, if an investigator wants to prove that a new drug is better than the standard one for treating a certain type of tumor, the null hypothesis would be:

H_0: The new drug has a similar (not better and not worse) frequency of adverse effects compared with the standard one for treating castrate-resistant prostate cancer.

2.6.2 Alternative Hypothesis

The alternative hypothesis, as the name suggests, represents the opposite idea of the null hypothesis, being shown when the H_0 is rejected. It is usually represented as H_a or H_1 and it is intended to state the nature of the difference, if one truly exists. For this reason, the alternative hypothesis must ideally be established right after the null hypothesis has been defined, consequently allowing the researcher to have the study's intent clear and well developed.

Based on the example of the superiority trial mentioned above, the corresponding alternative hypothesis could be:

H_a: *The new drug has a lower frequency of adverse effects than the standard one for treating prostate cancer.*

or

H_a: *The new drug has a greater frequency of adverse effects than the standard one for treating prostate cancer.*

Alternative hypotheses in superiority trials in oncology are generally two-tailed, which means the study evaluates whether the experimental therapy is *different from*, rather than only *better* than the control treatment. That is why a researcher would employ the statistical test used to compare the endpoint results in both arms in the above example, investigating whether the new drug offers more or fewer adverse events than the standard treatment, not only fewer adverse events.

2.6.3 Non-inferiority × Superiority × Equivalence Questions

Another criterion that might alter the course of defining the research question/hypothesis is the idea of whether an intervention is non-inferior to or better than the other, or equivalent.

- *Non-inferiority*: A non-inferiority question aims to demonstrate whether a new therapy is not clinically worse than the standard one by more than a pre-specified boundary, a pre-specified non-inferiority margin [7]. Investigators usually choose this type of question rather than a superiority trial, which is the most conventional, when the purpose is to approve more convenient, less toxic, or cheaper interventions.
- *Superiority*: This question investigates whether a new therapy has better efficacy when compared with that of the standard drug or a placebo. In this type of trial, it is fundamental to consider the required number of subjects and the risk of type I and II errors as, to prove superiority, a large sample size is often vital for the potential rejection of the null hypothesis. In phase III registration superiority cancer trials, the test of the hypothesis is generally two-sided, i.e., investigators evaluate whether the experimental intervention is better or worse than the control treatment.
- *Equivalence*: Finally, this hypothesis asserts that the new therapy is equivalent to the conventional treatment. Therefore, it intends to establish that the two therapies being compared have similar effects. This type of hypothesis should be used when the new treatment is simple, less expensive, or has fewer side effects than the standard therapy, even if it does not reflect a greater therapeutic effect than the standard therapy. Unfortunately, this type of hypothesis is not a very common scenario in the oncology field.

2.7 Summary

As discussed in this chapter, it is essential to spend a significant amount of time on thinking about and developing a research question. Usually, this is an interactive process in which there should be several drafts of the research question. A suggestion for new investigators is to test their questions with more experienced researchers. After testing and discussing the research question, the next step is to refine the approach and protocol design (as discussed in chapters 10, 11 and 12). The thought process for the formulation of a research question is crucial, because appropriate methodology and robust analyses cannot counteract or amend an inappropriate research question.

References

1. Hulley S, Cummings S, Browner W, et al. Designing clinical research. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2007.
2. Halvorsen A, Kristensen G, Embleton A, Adusei C, Barretina-Ginesta M, Beale P, Helland Å. Evaluation of prognostic and predictive significance of circulating micro RNAs in ovarian cancer patients. Dis Markers. 2017;2017:3098542.
3. Wang Y, Kong X, Guo Y, Wang R, Ma W. Continuous dose-intense temozolomide and cisplatin in recurrent glioblastoma patients. Pany S, ed. Medicine. 2017;96(10):e6261. https://doi.org/10.1097/MD.0000000000006261.
4. Middleton G, et al. Vandetanib plus gemcitabine versus placebo plus gemcitabine in locally advanced or metastatic pancreatic carcinoma (ViP): a prospective, randomised, double-blind, multicentre phase 2 trial. Lancet Oncol. 2017;18(4):486–99. https://doi.org/10.1016/S1470-2045(17)30084-0.
5. Shipley WU, et al. Radiation with or without antiandrogen therapy in recurrent prostate cancer. N Engl J Med. 2017;376(5):417–28. https://doi.org/10.1056/NEJMoa1607529.
6. Guan Y, Liu S, Wang H-Y, et al. Long-term outcomes of a phase II randomized controlled trial comparing intensity-modulated radiotherapy with or without weekly cisplatin for the treatment of locally recurrent nasopharyngeal carcinoma. Chin J Cancer. 2016;35:20. https://doi.org/10.1186/s40880-016-0081-7.
7. Riechelmann RP, Alex A, Cruz L, Bariani GM, Hoff PM. Non-inferiority cancer clinical trials: scope and purposes underlying their design. Ann Oncol. 2013;24(7):1942–7. https://doi.org/10.1093/annonc/mdt073.

Types of Variables and Distributions

3

Gregory R. Pond and Samantha-Jo Caetano

3.1 Introduction

One of the first things necessary for researchers to understand when performing a research study is to identify their research question and to explicitly state the main goals of their study. These are often referred to as the research objectives. These research objectives describe how that research will improve our understanding of the system under study. These research objectives may be identified as of primary, secondary, tertiary, or even exploratory, interest for investigators. To perform this research, investigators typically observe the effects of an intervention on a small group of individuals, called a sample, who are often defined by certain characteristics. Ideally one studies the effects of the intervention on the entire population, but it is generally unfeasible to measure every individual in a population at any time, and as a result, the true effect on the population will never be known with certainty. Hence, the effect of the intervention on the sample of individuals is important in that it represents how the intervention might affect the larger population of individuals who have similar features. This representation is formalized by using statistical inference. Therefore, on the basis of the study sample, investigators will make inferences on how the new treatment would work if given to all patients in the study population. To make the most accurate inferences, the study sample should be randomly selected from the larger population. This means that every individual in the population has an equal chance of being included in the smaller study sample. In this way, the sample is representative of the study population, and this ensures valid statistical inference.

G. R. Pond, Ph.D., P.Stat. (✉) • S.-J. Caetano, M.Sc.
McMaster University, Hamilton, ON, Canada
e-mail: gpond@mcmaster.ca

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_3

After identifying the research goals, investigators must define how they will measure the effect of the intervention so that they can interpret their research. Unfortunately, objective measurements are not always straightforward to define and an incorrect definition could negatively affect the ability to interpret study results or to infer these results to the general population. Therefore it is important for investigators to have a thorough understanding of how items can be measured and how these measurements can impact the optimal design and interpretation of a study. This chapter will therefore summarize the different ways that items can be measured, and the implications of these choices in study design and interpretation.

3.2 Variables

3.2.1 Variables and Parameters

To first understand how one can measure the effects of an intervention, there are some statistical terms which must be defined: statistics, variables, and parameters. For statisticians, a variable is a measured value which can vary from individual to individual [1]. In contrast, a parameter is a quantity which helps define a theoretical model, which often can be used to describe the population [1]. Lastly, a statistic is a quantifiable measurement from the study sample that is used to summarize a variable.

Assume, for example, that investigators are interested in studying the effect of a new drug on patients with a particular type of cancer, say stage IV pancreatic cancer. They give the new drug to a sample of patients with stage IV pancreatic cancer and observe the effects. Another group of investigators may examine the effects of the same new drug on a different sample of patients with stage IV pancreatic cancer. Since the two samples of patients are different, they will also differ in terms of the patient characteristics, notably the patient ages, sexes, weights, and disease histology. These characteristics are types of variables, and each one of these variables can be measured and described using a particular statistic. One might describe the average patient age or weight in each sample using statistics such as the mean, standard deviation, median, minimum, or maximum. A proportion and a ratio are statistics which could be used to describe the patient sexes or disease histologies within each sample. If we assume that both samples are randomly selected from the population, we can then try to infer information about the true population parameters of interest, such as the population mean age or the proportion of stage IV pancreatic patients who are female. Therefore, ideally, one selects a statistic which gives a representation of the sample. For example, the mean age of a sample gives information about what the ‘average’ age of the people is; in contrast, the mode would not be all that useful in describing the age of people. Alternatively, if one is interested in describing the overall survival time, or time to death, amongst patients with stage IV pancreatic cancer, the median may be a better descriptor of the average survival time than the mean. This is because a few people with stage IV may live a long time, potentially a few years, but the majority of people will only live for a few months.

Having one or two people who live a long time will greatly affect the mean, but will marginally affect the median. Hence, the median may better represent the time to death in the majority of patients. Ultimately, measuring the median (statistic) time to death (variable) would then allow investigators to infer information on the population median (parameter), which is never known with certainty.

3.2.2 Types of Variables

For a small number of subjects, it is quite simple to describe a group of individuals based on their specific variable values so that others can interpret them. For instance, if a 70-year-old, 200-lb male with stage I adenocarcinoma pancreatic cancer and a 64-year-old, 150-lb female with stage II undifferentiated pancreatic carcinoma both received the novel treatment, one can measure their survival time and easily interpret the results. If the patients' survival times were 32 and 38 months, one can directly interpret whether these results are similar to, better than, or worse than what one might expect for comparable individuals. However, it is much more difficult to describe and understand these relationships when there is a large number of individuals in a study. Descriptive statistics are a useful way to summarize this information, and provide an understanding of the variables and the effects of the treatment. As discussed above, the specific statistic which is used depends on the nature of the variable. In general, variables can be grouped into one of four distinct types: categorical, continuous, ordinal, and time-to-event/survival [2].

Categorical variables are measured values which can be classified into two or more distinct groups, and individuals can belong to one, and only one, category at a time. Another feature of categorical variables is that there is no natural way to order the groups. For example, cancer histology is a categorical variable. One investigator might present information on each category of cancer histology, preferring to describe adenocarcinomas first, followed by squamous cell cancers, and then by carcinoid. Another investigator may present carcinoid information first, followed by adenocarcinoma, and then squamous cell cancer. Again, there is no natural or biological reason why one ordering might be preferred over the second ordering. Specific cancers can belong to any one histological subtype, but cannot be in multiple subtypes at the same time. A special type of categorical variable occurs when there are only two groups, such as being alive or dead. When there are only two categories, it is called a dichotomous variable.

However, variables which can be divided into distinct groups that include a natural ordering are called ordinal variables. The major distinguishing characteristic between an ordinal variable and a categorical variable is the presence of a natural or biological ordering. For example, the stage of cancer disease is an ordinal variable, whereby stage II patients have generally worse disease than stage I patients. Similarly, stage IV patients have worse disease than stage III patients. For ordinal variables, the gradient between different groups does not have to be constant; for example, the expected survival for stage II patients may not be much worse than that for patients with stage I disease; however, there may be a very large difference between those having stage IV and those having stage III disease. Similarly, toxicity may be graded using an ordinal scale, such as the National Cancer Institute Common Terminology Criteria for Adverse Events (NCI-CTCAE) scale, where 0 represents no toxicity, 1 = mild, 2 = moderate, 3 = severe, 4 = life-threatening, and 5 = death (Cancer Therapy Evaluation Program; CTEP). Similar to categorical variables, individuals can be in one group and only one group at a time, and cannot be in between groups. For instance, a patient cannot have a stage 3.5 cancer.

Variables which can be measured using a numerical scale are called continuous variables. Continuous variables can be measured in smaller intervals, where the preciseness of the measurement is restricted to the preciseness of the tool doing the measurement. Additionally, each unit change in a continuous variable is constant from one unit to another. For example, age is a continuous variable. Age can be measured in years, months, days, seconds, milliseconds, etc. The level of preciseness depends solely on the measurement tool, and the level of accuracy required. The use of years to describe the age of an individual in most circumstances is for simplicity purposes. In addition, the difference of one unit, or year, is constant, regardless of the age of an individual. Two people who are born on the same day, but 1 year apart, will be 1 year apart in age regardless of whether they are 2 and 3 years old, or 92 and 93 years old. Another continuous variable is a person's hemoglobin count, which can be reported in terms of units g/dL. Most measurements will report hemoglobin to the tenth of a unit, and although smaller and smaller units may be calculated, they tend not to be of great use (i.e., one could differentiate between an individual with hemoglobin of 13.1001 g/dL and another with 13.1002 g/dL; however, for practicality, there is no functional difference between these values) and most laboratory tests are not precise enough to distinguish such small differences.

The fourth type of variable commonly used is survival, or time-to-event variables. Time-to-event variables are continuous measures that contain missing information since some data is unknown or incomplete at the time of any analysis. For instance, overall survival, or time-to-death, can be measured as a continuous value such as the number of days until death. If everyone in the study has died at the time of analysis, then the variable is simply a continuous measure. However, it could take a long time for every individual to die, and scientific advancement will not occur if researchers must wait for every individual to die in their study before reporting on their research. One cannot simply exclude all individuals who have not died, since they contribute some partial information. These individuals are considered 'censored' and the time to censoring is included in survival analyses. Special types of statistical analyses are required to evaluate time-to-event or survival data. In fact, analysis of this type of data is so important that an entire Sect. 3.2.3 of this book is devoted to the analysis of survival type data (Chap. 7).

3.2.3 Independent vs. Dependent Variables

With an understanding of the different types of variables, investigators must also know what the purpose of each variable would be within a given study. The outcome variable is the main measure which will be used to objectively allow researchers to decide whether their study result is positive or negative. The outcome measure is also called the dependent variable, since the outcome variable will depend on what the sample looks like. The measured values of the data point which can influence the outcome are called independent or predictor variables. Predictor variables describe traits of the individuals which may influence the outcome, or may help investigators understand the characteristics of the people included in the study sample, which will allow others to make an inference about the characteristics of the population.

For example, a group of individuals which has a greater percentage of patients with early-stage cancer might be expected to live longer, on average, than a group of individuals who have a higher percentage of patients with late-stage disease. In this manner, the stage of disease is considered the predictor variable, and it is being measured by the statistic: percentage or proportion. The outcome variable in this example is then the survival time. To test the effect of an intervention, one might measure the differences in survival time for patients who receive an intervention, and compare this with the survival time of patients who did not receive the intervention. To do a valid comparison, one might have to account for, or adjust for, the predictor variables. The investigators may therefore adjust for the predictor variable 'stage', since the study is interested in the effect of the intervention, and not the effect of disease stage, which may be influencing the outcome. Notably, variables can be predictor variables in some studies, but outcome variables in other studies. A research study looking at prevention of cancer might be looking at whether an intervention can prevent the occurrence of high-stage prostate cancer among men who were initially diagnosed with Gleason score 6 prostate cancer. Disease stage is therefore an outcome variable in this study, and not a predictor variable as described in the previous example.

It is imperative that investigators define and appropriately select variables for analysis in their research studies. Outcome variables must be selected to represent clinically relevant information so that the study is impactful [3]. Studies in cancer often use overall survival as the outcome variable of interest; however, overall survival may not be a relevant outcome in a study of palliative care patients with cancer, or in studies of patients receiving adjuvant therapy for ductal carcinoma in situ. Additionally, investigators must select outcomes which could plausibly be impacted by the intervention under study. This is why studies of radiation therapy interventions commonly use time to local recurrence as a main outcome, or why a quality-of-life measure might be used when investigating a palliative therapy. It is important to simultaneously recognize the potential clinical importance of an outcome. While a quality-of-life measure is important in a palliative care setting, improvements in quality of life may not be viewed as important if they come with a substantial reduction in overall survival.

It is also important for investigators to properly and clearly define the variables, including how they will be measured. When doing this, the investigators must be very specific. Many common terms have different definitions depending on the study; for instance, it is not uncommon for progression-free survival to have multiple definitions in different studies [4]. By being precise and consistent with their definition, investigators ensure that their research study can be replicated. For example, some investigators might include contralateral breast cancer, distant breast cancer, or development of a new primary breast cancer as a progression event when the outcome is progression-free survival, while others may not. Further, some investigators may want to include only deaths that are definitely related to breast cancer, while others may want to include death due to any cause. Standardized definitions should be used whenever possible [5]. Finally, one must be clear on how the variables are measured. Are the investigators interested in the median overall survival, the 1-year overall survival rate, or overall survival across all time points? Defining these outcomes in the study planning phase will help to avoid long-term problems when the time for analysis comes.

3.3 Distributions of Variables

Generally, researchers study a sample of individuals to perform a scientific experiment, and then try to infer how these results relate to the entire population. Statistical theory tells statisticians how the population values relate to the sample values. This statistical theory is based on knowledge of statistical distribution, which is a mathematical function describing the range of possible values and the likelihood of observing each possible value. There are infinite numbers of possible distributions, but some of the more commonly used distributions include the normal (Gaussian), exponential, Poisson, and binomial distributions. Each of these distributions can be completely specified using one or two parameter measures. For instance, the normal distribution is completely specified if one knows the population mean and standard deviation, whereas the exponential distribution is completely specified by the beta (or lambda) parameter.

To better understand the concept of distributions, assume one has a fair coin, which means the probability that the coin lands on heads is 0.5. As an experiment, it might be of interest to find out how many times the fair coin would land with heads facing up, if the coin was flipped 10 times. It is most likely that the coin would land on heads 5 times; however, it is quite possible that, for a given experiment, the coin might land on heads a different number of times. In a single experiment, an investigator might flip 4 heads, or 3, and so on. The probability of each scenario can be calculated exactly, and these values are shown in Table 3.1. One can see that even though it is unlikely, it is still possible that one might flip 10 consecutive heads even with a fair coin, as the probability is 0.001 (i.e., if 1000 people flipped 10 coins, it would be expected that one person would flip 10 consecutive heads, and another person would flip 10 consecutive tails!). The range of possible values can be plotted, and this forms a distribution, which is shown in Fig. 3.1. In fact, this is a binomial distribution with $n = 10$ and parameter $p = 0.5$.

Table 3.1 Range of possible number of 'heads', and the associated probabilities when flipping a fair coin 10 times

Number of heads	0	1	2	3	4	5	6	7	8	9	10
Probability	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Again, there are infinite numbers of distributions, but only a select few are commonly used. The most well-known distribution is the normal distribution, also called the Gaussian distribution [6]. A normal distribution is perfectly symmetric around the mean parameter. Roughly two-thirds of all data points in a normal distribution will be within ±1 standard deviation of the mean, and 95% of data points will be within ±2 standard deviations of the mean. Theoretically, there is no upper or lower limit of a normally distributed random variable. The normal distribution is a continuous distribution, which means that it can be used to represent continuous variables. The standard normal distribution has mean 0 and standard deviation equal to 1 and is shown in Fig. 3.2.

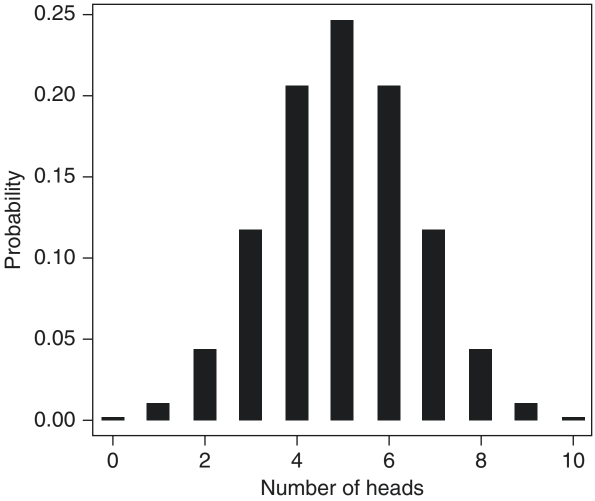


Fig. 3.1 Example of binomial distribution with $n = 10$ and parameter $p = 0.5$

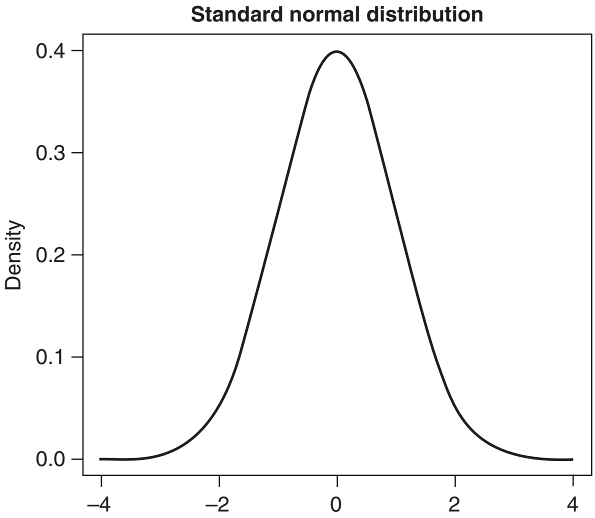


Fig. 3.2 Standard normal distribution

The normal distribution is very valuable in statistics, owing to the central limit theorem (CLT) [7]. Specifically, the CLT states that for relatively large sample sizes the sampling distribution of the sample mean for any random variable will approach a normal distribution. Although there is no absolute rule, the CLT is appropriate to use for 'large' sample sizes, which is sometimes interpreted as 30 (or 50) or more. Further to this, the distribution of the mean of any sample, if sufficiently large, will be approximately normally distributed, with the mean of the sample mean equal to the mean of the population, and the standard deviation of the sample mean equal to the standard deviation of the population mean divided by the square root of the sample size. So, one can take a random sample of 200 patients and calculate the mean age. Another person can take a similar random sample of 200 patients and calculate the mean age of the second sample. A third person calculates the mean age of a third random sample, and so on. As the sample size increases, the distribution of means will look more and more like a normal distribution. What makes the CLT so powerful is that it does not matter what the shape of the population distribution is, making it applicable in all situations with relatively large sample sizes. Even more impressive, the CLT still applies even if the underlying distribution is not continuous. For instance, one might notice that the distribution presented in Fig. 3.1, which is a binomial distribution, looks a lot like the normal distribution presented in Fig. 3.2. Hence, even though the binomial distribution can be used for dichotomous outcomes, one can estimate the parameter p, the proportion, by invoking the CLT. The distribution of p is therefore a normal distribution, and inferences can be made from the known distribution.

The CLT lays the foundation for many common statistical techniques, such as hypothesis testing, significance testing, and confidence interval estimation. Since one knows that the sampling distribution for the sample mean of any population distribution will be approximately normal, one can approximate the population standard deviation from the standard deviation of a given sample, and make inferences about the population mean (which is the only information that is unknown). If one makes an assumption about the population mean, or creates a null hypothesis, then one can examine the likelihood that the observed data came from that particular distribution. How consistent (or inconsistent) the sample data is with the assumption can be quantified, and can be represented by the commonly used p-value. Further, if one makes two competing assumptions, say the null and alternative hypotheses, then one can test which is more plausible, given the data. This is the basis for hypothesis testing. Since the mean is a good representation of the 'average' in many situations, one can then evaluate the effects of an intervention on the 'average' individual in a study, or one can compare the means between two different interventions, and test the hypothesis that they are equal.

Unfortunately, not all experiments are as simple as comparing the differences between two means. Sometimes, the approximation to the normal distribution is not sufficiently precise, and investigators may want more accurate estimates of the p-value. In other studies, the mean may not be the best statistical measure to use to represent the population average. For example, when discussing time-to-death, or overall survival, the distribution is askew and the mean is no longer in the 'center' of the distribution. Figure 3.3 shows an example distribution that represents survival time for patients with a deadly form of cancer; this is a highly skewed distribution. Since one or two patients who live much longer than everyone else could cause the mean survival time to be much higher than the survival time of most people, the mean may not be a good descriptor of the ‘average’, or ‘center’, and other statistics, such as the median, may better represent the ‘average’. Although the CLT still holds, it may be better to compare the median survival times rather than the mean survival times when evaluating different treatments. Therefore it is important to know other distributions.

The exponential distribution is another common distribution, which is often used to model time between events. For example, the exponential distribution is often used to model survival times for patients with serious diseases such as cancer. There are infinite numbers of exponential distributions, each specified by a single parameter, λ. The mean of each exponential distribution is $1/\lambda$ and the variance is $\lambda^{\frac{1}{2}}$. Some examples are illustrated in Fig. 3.4. One of the important properties of the exponential distribution is the ‘memoryless’ property, which states that the probability of an event occurring in the next time period, i.e., $(t, t + 1)$, provided that the event has not occurred prior to time t, is the same as the probability of the event occurring in the first time period, (0,1). For example, if a patient’s survival time follows an exponential distribution, and that patient is alive after 3 years, the probability of that patient dying in the next year (i.e., between year 3 and year 4) is the same probability that the patient had of dying in the first year (i.e., between year 0 and 1). Mathematically, this is $P(T > t + a|T > t) = P(T > a)$.

If the time between events, or deaths, follows an exponential distribution, then it is interesting to note that the number of events within a given time period follows the Poisson distribution with parameter λ. The Poisson distribution represents

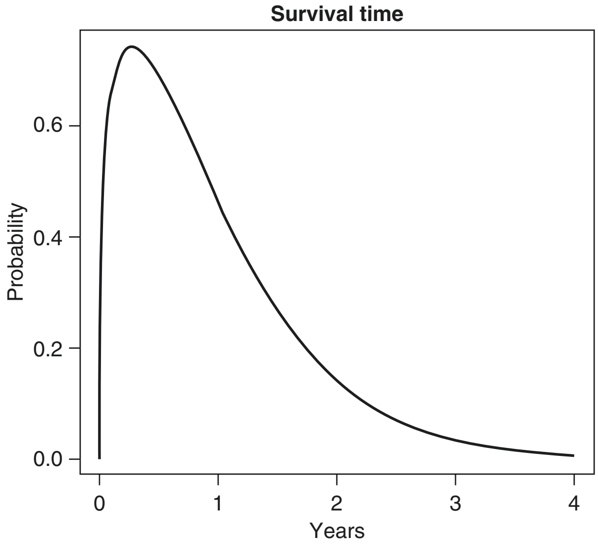
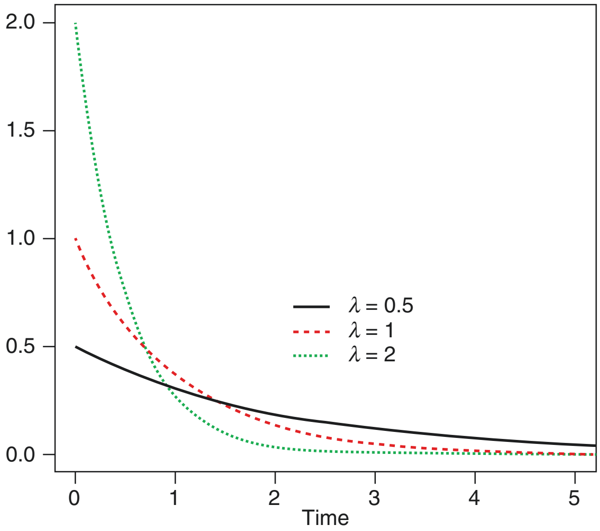


Fig. 3.3 Example of hypothetical survival times of patients

Fig. 3.4 Example of exponential distributions with parameter $\lambda = 0.5$, 1, and 2



counts, or the number of events in a given interval of time period (or space). The mean number of events in a given time period is λt, where t is the fixed interval of time, and the variance is also λt. For example, hospitals may want to know how many patients died in their intensive care unit within successive months. If the time between deaths follows an exponential distribution with parameter $\lambda = 2$, then the mean time between deaths is $1/\lambda =½$ or 0.5 months. Similarly, the mean number of deaths per month is $\lambda t = 2(1) = 2$.

3.4 Confounding Variables

Another issue that investigators must be aware of is the presence of confounding variables. A variable is confounding if it is related to both the predictor and outcome variables, and the relationship between the predictor and outcome variables is affected because of the effect of the confounding variable [8]. One example of a confounding variable is that, in retrospective studies, patients who are treated with chemotherapy may appear to have worse outcomes than those treated without chemotherapy. However, the fact is that patients who have more aggressive disease are more likely to be treated with more aggressive treatments such as chemotherapy. Hence, disease status is a confounding variable in this example. In research studies, the effect of confounding variables may mask or distort the true effect of an intervention. Sometimes an effect may be exaggerated, and other times it may be minimized, owing to confounding variables. In either situation, the confounding variable is biasing the results.

In an ideal situation, a scientific experiment is performed in conditions that control all possible effects except for the one under study, usually the intervention. However, when dealing with human investigations, one cannot control everything. In clinical studies, there will always be confounding variables, such as disease stage, number of comorbidities, or the sociodemographic characteristics of patients. Fortunately for statisticians, the impact of confounding variables can be minimized through randomization and random selection, provided the sample sizes are large enough [9]. As the sample size increases, statistics tend to become closer and closer to the true population parameter value. The variability around sample statistics, and as a result, the sample distributions, decreases and tends toward 0 as the sample size increases. Similarly, in large samples, by randomly allocating patients to one of two treatment arms, the effect of confounding variables is minimized, since the distribution of the confounding variables is likely balanced between the treatment arms, which decreases the variability between the groups. Therefore any impact of confounding variables between samples (treatment arms) will be small. This is true for both measurable and known variables, as well as non-measurable or unknown variables.

Unfortunately, many studies do not have large sample sizes. Cancer clinical trials, for instance, are often performed with a couple of hundred patients enrolled, or even fewer. Therefore, with a moderate or small sample size, large differences may remain between sample groups, just owing to chance alone. This is compounded by the fact that there are often multiple potential confounding variables (for example, histology, stage, age, prior treatment, and comorbidities may all impact patient survival) and even the occurrence of just one confounder may grossly impact the results. Therefore, to control as much as possible and make the groups as similar as possible for comparison, investigators often create balance artificially through quasi-random methods of allocation. These methods can improve balance between arms, but they do come at a cost against pure randomization.

Common quasi-random methods of allocation include dynamic allocation, minimization, stratified random sampling, or permuted block designs [10]. In permuted block randomization, researchers identify, prior to the study, those variables they believe pose the greatest risk of confounding. Researchers stratify patients into risk groups (e.g., old versus young patients, high stage versus low stage, no comorbidities versus at least one comorbidity, etc.). The individual patients are grouped into strata, or groups, by the stratification factor. Within each stratification factor, blocks of equal size are created which dictate how patients will be allocated to each treatment arm. For instance, there are six permutations, of size four, when two patients receive intervention A, and two receive intervention B: AABB, ABBA, ABAB, BBAA, BAAB, and BABA. A block is randomly selected from these six possible permutations, and patients are allocated to arm A or arm B in order, as dictated by the selected block. Hence, after every group of four patients, there will be two who receive arm A and two who receive arm B. Separate block patterns are created within each stratum, so that at the end of the trial, there will be a nearly equal number of high-risk and low-risk patients receiving each of the two treatments.

Another method, called dynamic allocation, or often misnamed minimization, uses the information of patients already recruited to a study, and then preferentially allocates the next patient to the arm which best balances the confounding factors between the two arms [11]. For example, in a prostate cancer trial, one might have strata for age, prostate-specific antigen (PSA), and number of comorbidities. If the next patient to be enrolled in the trial is young, has low PSA, and a high number of comorbidities, then dynamic allocation methods are used to assess how many young patients, how many patients with low PSA, and how many patients with a high number of comorbidities, are in each treatment arm, respectively. The patient of interest would then be allocated to the treatment arm which minimizes any imbalance. A random component could be added to ensure that the model is not completely deterministic. There are numerous variations of these types of designs, including a biased coin, stratified sampling, and so on. Minimization is, in fact, a very specific type of dynamic allocation method [12].

Regardless of the method selected for balancing confounding variables, if any method is selected at all, it is important to account for potential confounders in the statistical analysis. This is usually performed using regression analyses. The type of regression analysis to use depends on the outcome variable. For categorical variables, logistic regression is used, whereas for continuous variables, linear regression is most common. Cox proportional hazards regression is used when the outcome is a time-to-event outcome. By performing a regression analysis which adjusts for potential confounding variables, any impact of confounding is minimized.

Conclusion

After the formulation of the research question, the next step is the determination of the study primary endpoint. To evaluate the study endpoint, investigators need to be able to describe what the endpoint might look like, which requires an understanding of the different types of variables. The type of variable selected will then affect the types of study methods, statistical analyses, sample size, and ultimately, study budget and design. Understanding statistical distributions is imperative for planning and interpreting statistical analyses. Given the importance of variables and statistical distributions in study design, we strongly recommend that researchers consult a statistician when planning these elements. We also recommend the following text: http://www.biostathandbook.com/confounding.html.

References

1. Altman DG. Statistics notes: variables and parameters. Br Med J. 1999;318:1667. https://doi.org/10.1136/bmj.318.7199.1667.
2. Larson MG. Descriptive statistics and graphical displays. Circulation. 2006;114:76–81. https://doi.org/10.1161/CIRCULATIONAHA.105.584474.
3. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? J Clin Oncol. 2012;30(10):1030–3. https://doi.org/10.1200/JCO.2011.38.7571.
4. Saad ED, Katz A. Progression-free survival and time to progression as primary end points in advanced breast cancer: often used, sometimes loosely defined. Ann Oncol. 2008;20(3):460–4. https://doi.org/10.1093/annonc/mdn670.
5. Gourgou-Bourgade S, Cameron D, Poortmans P, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-Event Endpoints in Cancer Trials). Ann Oncol. 2015;26(5):873–9. https://doi.org/10.1093/annonc/mdv106.

6. Altman DG, Bland JM. Statistics notes: the normal distribution. Br Med J. 1995;310:298. https://doi.org/10.1136/bmj.310.6975.298.
7. Shang Y. A note on the central limit theorem for dependent random variables. ISRN Probabil Stat. 2012;2012:192427. https://doi.org/10.5402/2012/192427.
8. McDonald JH. Handbook of biological statistics. 3rd ed. Baltimore, MD: Sparky House Publishing; 2014. p. 24–8.
9. Lachin JM. Statistical properties of randomization in clinical trials. Control Clin Trials. 1988;9(4):289–311. https://doi.org/10.1016/0197-2456(88)90045-1.
10. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet. 1990;335:149–53. https://doi.org/10.1016/0140-6736(90)90014-V.
11. Pond GR. Statistical issues in the use of dynamic allocation methods for balancing baseline covariates. Br J Cancer. 2011;104(11):1711–5. https://doi.org/10.1038/bjc.2011.157.
12. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975;31(1):103–15. https://doi.org/10.2307/2529712.

Testing Measures of Associations

4

Bruno S. Paolino, Raphael L. C. Araújo, and David Bristol

4.1 Introduction

The choice of tests of measures of association is crucial to any research plan and has to be one of the first decisions in a study methodology. The decision about which test to use is made according to the hypothesis addressed in the research project, and is closely related to the study primary endpoint and methodology. The selection of a test should follow the hypothesis and study endpoints in a way that determines how and which data will be collected, the way to measure and define them, and also the obvious strictness to avoid biases. The selection of an inappropriate test might lead to either false-positive or false-negative associations, and may compromise the internal validity of the study, leading to wrong interpretations, and consequent waste of time and resources. The selection of the correct test should be based on an understanding of the test of measures of association, recognizing their role and limitations. Although a statistician should be consulted to support the decision on the tests, it is strongly recommended that every researcher should have basic knowledge of this topic.

A statistical association means any statistical relationship between two or more groups of variables, regardless of whether it is causal or not. A measure of association attempts to estimate the strength of any association between two variables. Tests of association concern the detection and calculation of the effect size of a given association, considering the probability of the event of interest, whether it

B. S. Paolino, M.D., Ph.D. (✉)
Department of Cardiology, State University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

R. L. C. Araújo, M.D., Ph.D.
Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos Cancer Hospital, Barretos, SP, Brazil

D. Bristol, Ph.D.
Independent Consultant, Wiston-Salem, NC, USA

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_4

happened by chance or not. Considering the null hypothesis as the absence of association between two variables, tests of association can rule out the null hypothesis, when it demonstrates that the groups of variables are different, or fail to reject the null hypothesis, when the numeric difference between the groups is not sufficiently large. However, both the rejection and the assumption of a null hypothesis can happen, representing a false-positive and a false-negative error, respectively.

The false-positive error, also called a type I error, represents the probability of rejecting the null hypothesis when, in fact, it is true, based on the low degree of significance. The probability of a type I error is typically referred to as alpha (α). The *p*-value represents the probability that the result of a test was achieved by chance and not based on true association. A *p*-value less than the specified *α* is considered statistically significant. The most common choice for α is 0.05. In general, readers of clinical research overestimate the importance of the *p*-value. The *p*-value can demonstrate whether two variables are statistically associated or not, but that does not necessarily mean that the association is clinically relevant, for instance. Moreover, a smaller *p*-value does not necessarily signify a more important clinical effect [1]. To better interpret the results, *p*-values and confidence intervals should be presented together, since the confidence interval might hint at the effect size as well and the *p*-value alone cannot indicate the importance of a finding [2].

At the same time, the type II error, or false-negative, can happen when the null hypothesis was not rejected because the differences were not sufficiently large to be detected with the selected sample size or test. Type II errors will be addressed in depth in the “Sample Size Calculation” chapter (Chap. 5).

In this chapter, we will discuss the most commonly used tests of measure of association in clinical oncology research, their advantages and disadvantages, when to use each test, and how to interpret the results accurately. Note that, in this chapter, we will discuss some basic types of variables and dispersions, which are key to understanding tests; however, these types will be discussed in depth in a separate chapter (Types of Variables and Distributions; Chap. 3) Statistical tests utilized to analyze time-to-event variables will also be covered in a separate chapter, entitled “Survival Analysis” (Chap. 7).

4.2 How to Choose Your Test of Measure of Association

4.2.1 Basic Concepts

The choice of the optimal test depends on the sample size, the number of groups, the nature of the variable (categorical or continuous), and type of dispersion—parametric or nonparametric. There are various software packages that are worthwhile tools to run the tests, but the decision about the best test for your hypothesis and data relies on you, the investigator. Therefore, we strongly recommend the support of biostatisticians at the beginning of study planning, before the start of data collection. Since we are looking into the understanding of concepts of tests of measure of association, mathematical algorithms will not be the focus of this chapter.

As the types of variables and their dispersion have been discussed previously (Chap. 3), in this chapter we will give an overview of types of variables and discuss deeper examples of tests of measures. Briefly, there are two types of variables: quantitative and categorical ones. A quantitative variable can be described as a mathematical value. For instance, if you are looking at the number of recurrent tumor sites after a chemotherapy regimen, or their size in centimeters, you are using a quantitative variable. Note that in the first example, the number of lesions that recurred, data are always expressed in whole numbers. In the second one, the size of the tumor in the recurrence, the tumor size can be a number such as 0.6 or 0.8 cm, or any other number depending on the precision of the assessment. These kinds of quantitative variables are called, respectively, discrete and continuous variables. Discrete variables might be analyzed as continuous when using a test to measure an association, as an approximation.

Categorical variables are not expressed in mathematical values since there is no intrinsic quantitative value associated with them. They can be used as binomial (yes/no) or ordinal (scores) variables. For instance, gender and vital status of the subjects in a sample are binomial variables, given that the possibilities—man versus woman or alive versus dead, respectively—are not quantitative but qualitative. Qualitative variables are expressed in frequencies or proportions.

One important advantage of quantitative variables is that they can be transformed into categorical variables. For instance, if you have measures of creatinine, you can point out a cut-off (creatinine 1.5 mg/dL, for instance) and transform this continuous variable in a yes/no decision, such as kidney dysfunction versus normal kidney function. Note, however, that the cut-off needs to be meaningful to the research and valid for the scientific community—and, obviously, determined *before* the collection of data. The disadvantage of a categorical variable is the loss of power in tests to measure association, compared with a continuous variable, because slightly different characteristics can be interpreted as similar, potentially losing power in the statistical analyses. In the example above, patients with creatinine levels of 1.6 mg/dL will be classified as having kidney dysfunction and lumped together with patients needing dialysis, i.e., those with creatinine levels of 10.0 mg/dL. Thus, categorical variables constructed from continuous variables often require larger sample sizes to achieve the same control of the probability of a type II error. Even if you want to show the variables in descriptive forms, the use of a test for continuous variables and data transformation after the final results is a good strategy to protect data against loss of power.

Ordinal variables, even when expressed as numbers, must not be interpreted as quantitative. For instance, stages of a lung cancer, expressed as I, II, III, and IV, may give the idea of an ordinal scale—the higher the number, the more advanced is the tumor; however, this is not associated with the values because stage III is not three times more advanced compared with stage I. Because of this feature, statistical tests for categorical variables are used to measure associations between groups of ordinal variables.

Time-to-event analysis, commonly referred to as survival analysis, is common in clinical research and is widely used in oncology research. Briefly, it is a model for the assessment of time to the occurrence of a dichotomous event (i.e., alive versus dead). Thus, two variables are necessary for this analysis since we need the presence or not of the event of interest (yes or no) and a continuous variable to count follow-up (time). Survival analyses imply the use of specific analyses and such analyses and their inherent issues are discussed in depth in Chap. 7), as noted above.

4.2.2 Comparing Groups of Continuous Variables

The type of distribution is crucial for deciding the best way to compare two groups. Therefore, the first information you have to know about your data is which type of distribution it presents, whether it has normal (Gaussian) distribution or not, and, consequently, which kinds of central tendency measure and dispersion are more suitable—whether mean and standard deviation for parametric distributions or median and interquartile ranges for nonparametric distributions. Otherwise, mean and median could not be congruent, especially if there are outliers. In a hypothetical example of two graphs representing two different populations concerning weight distribution, depicted in Fig. 4.1, the curve "a" represents the normal distribution with coincident mean and median places. In contrast, the curve "b" is skewed to the left for two possibilities: small sample size, which has not yet assumed a normal distribution, or the presence of outliers that pushed the mean weight of the population to the right side. Since the median uses the most consistent half of the sample, the presence of the few outliers was not enough to move the median line to the right. Thus, the selection of the correct test based on variable distribution is imperative, and we recommend verifying, before the study, whether the distribution is normal.

The characteristic behavior of the variable can be tested to verify whether it follows the regular distribution, and some tests can be used to determine whether data

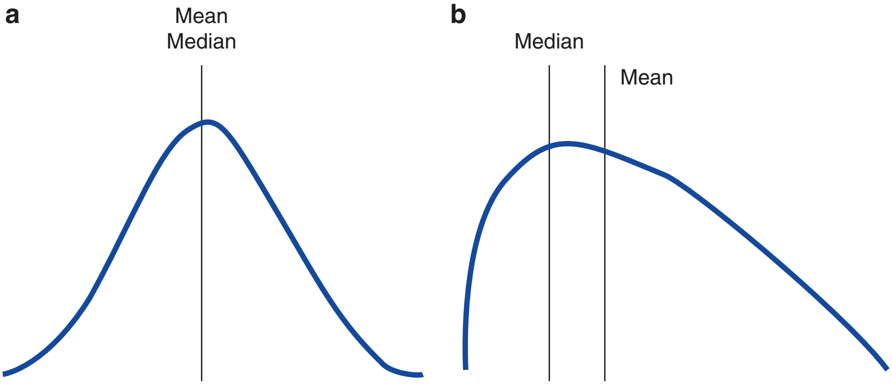


Fig. 4.1 Curves of variables distributions; (**a**) represents a normal distribution with overlapping of congruent mean and median; (**b**) represents a skewed distribution without equivalent mean and median

has normal distribution or not. One of the most popular and best-powered tests is the Shapiro-Wilk test [3], but other tests, such as the Kolmogorov-Smirnov, Anderson-Darling, and Lilliefors tests, can also be used to analyze whether samples have a normal presentation. In these tests, the null hypothesis demonstrates that the dispersion is normal and the rejection of the null hypothesis reflects data has non-normal distribution.

Regarding continuous variables, when the distribution is parametric (based on a normal distribution), data are compared using parametric tests, such as the *t*-test for comparison of two groups, and analysis of variance (ANOVA), which applies to three or more groups. The most commonly used test is the *t*-test, and it is based on mean and standard deviation. A simple example is to analyze the age distribution of patients who have undergone two different treatments or exposures. When the distribution is non-parametric (skewed, non-normal distribution), on the other hand or when the sample is small, some nonparametric tests must be performed to compare data, such as the Mann-Whitney (also called Wilcoxon rank-sum test) and Wilcoxon signed-rank test, which compare continuous variables between two groups, and the Kruskal-Wallis test, which compares three or more samples. Examples of Wilcoxon rank-sum tests are demonstrated in Table 4.2. The sample size also influences the distribution of data and therefore this size must be considered in the choice of the statistical test. The central limit theorem establishes that the distribution of the sample mean can be approximated by a normal distribution as the sample size increases—and a parametric test might be performed. In the same way, small samples of data often assume a non-normal distribution, even in a characteristically normally distributed variable. Importantly, one does not have to demonstrate a normal distribution for the assumption of a normal distribution to be true. If the sample size is small, a test based on the normal distribution can still be valid, although the power will be low.

4.2.2.1 Parametric Tests in Samples of Continuous Variables

The *t*-test, also called Student's *t*-test, is adequate to compare the means of two independent or dependent groups of continuous variables with normal distributions. Although the *t*-test can be used even when the data distribution is not normal, *t*-tests can be performed if the distribution is not excessively asymmetric and the samples are large. However, when the data do not follow a normal distribution, the *t*-test is not the appropriate choice and a nonparametric procedure may be more appropriate.

For independent samples—for instance, comparing two different groups' prostate-specific antigen (PSA) plasma levels after prostate cancer treatment— the *unpaired t*-test is suitable, as it compares both groups' data without matching subjects. However, when the groups are dependent—such as the PSA levels of a group of subjects at two different time points, such as before and after a new surgical procedure for prostate cancer—the *paired t*-test is the recommended test. While the unpaired *t*-test considers some variability of the two groups of independent subjects, in the paired *t*-test, each subject is compared with him/herself and no additional variance caused by the independence of data is included.

When more than two groups are to be compared, or there are more than two comparisons of the same groups, or there is more than one independent variable to compare, the ANOVA test is appropriate. There are many types of ANOVA that might be applied to different situations. Rejection of the null hypothesis means that there is a difference among the means, but does not indicate which comparisons are statistically different. To determine which individual comparisons are statistically significant, multiple comparison tests should be performed.

It is not possible to deal with all of the ANOVA types in this chapter; however, the types of and differences among ANOVA tests are summarized in Table 4.1. For example, if you want to compare the efficacy of three radiotherapy protocols for prostate cancer in a randomized trial and the outcome is the size of the tumor, one-way ANOVA compares all groups together, and might be used. However, if you have more than one independent variable—for instance, if you want to analyze the radiotherapy protocols and the use of testosterone blocker drugs or not in the previous example—two-way ANOVA is required. In two-way ANOVA, more than one *p*-value can be generated, as there is more than one independent variable. In two-way ANOVA, two levels of one independent variable and two levels of a second independent variable are analyzed [4]. In the example above, there would be: (1) a *p*-value for the comparison of radiotherapy protocols, independently of testosterone blocker use, (2) a *p*-value for the use of testosterone blockers, independently of radiotherapy protocols, and (3) a *p*-value for the interaction of both variables. When assessing three or more time points within the same cohort of subjects, repeated-measures ANOVA is required. For example, repeated-measures ANOVA was utilized in a cohort of 58 subjects with heart disease who underwent cardiac surgery; peripheral blood thyroid hormone measures were done at five different time points before, during, and after the procedure [5].

As mentioned above, when the null hypothesis is rejected using ANOVA, this means that there is a difference between the compared means, but multiple comparison tests are necessary to determine which comparisons are statistically significant. There are many types of multiple comparison tests and the main difference among them is the power to detect the difference and, consequently, the directly proportional risk of type I error. The risk of type I error in multiple comparisons is a real issue and should be considered, as the cumulative risk can exceed the desired 0.05 type I error, particularly when the number of groups or time points—and, consequently, the number of comparisons—increases. Therefore, it might be necessary to use the Bonferroni correction, which is an alpha level adjustment to control the final type-I error risk. The Bonferroni correction is often used, as it is easy to apply and interpret; it is calculated by dividing the *p*-value you desire (typically 0.05) by the number of comparisons [6].

The two most commonly used approaches to multiple comparisons are Tukey's honestly significant difference and the Bonferroni test, which are more conservative ones, in that they are less likely to detect differences among means, but protect the data against false-positive findings. Other tests with higher power are the Duncan multiple range test and Fisher's least significant difference test, i.e., they are more likely to find significant differences among means, but they are also more likely to provide false-positive results.

4.2.2.2 Nonparametric Tests in Samples of Continuous Variables

If one is comparing the time to complications after two surgical treatments (chemotherapy + surgery vs. surgery alone), probably the complications have peaks of incidence just after treatments and later at follow-up, causing a skewed distribution. The mean and the standard deviation may be inappropriate, because they will be influenced by the values of outliers, and so the median and the interquartile range are more suitable. For non-normal distribution variables or very small samples (when data often assumes non-normal distribution), nonparametric tests are preferentially performed, as they use ranks of the sample data instead of the specific values and, thus, do not require normal distribution. Another important indication for the use of nonparametric tests is when the data is described in ordinal variables. The main disadvantage of nonparametric tests is the loss of power to detect differences between groups, as these tests do not make any assumptions about the distribution of the original data. Therefore, when a nonparametric test is performed, compared with parametric tests, the number of subjects in each group must be increased to maintain the statistical power. The loss of power is particularly important in small samples, when the loss of power can reach 35%, compared with analog parametric tests. This power reduction is attenuated in large samples and then can be as low as 5% [7].

As shown in Table 4.1, for every parametric test, there is an analog non-parametric test with the same indications. The nonparametric equivalent of the unpaired *t*-test is the Mann-Whitney *U*-test. The analog of the paired *t*-test is the Wilcoxon signed-rank test. The latter is a very useful test because it allows comparisons of skewed curves. When it is necessary to compare more than two independent groups, the alternative to one-way ANOVA for skewed distribution samples is the Kruskal-Wallis rank test, and when there are more than two related groups, the Friedman two-way analysis of variance by rank test [8] is a powerful alternative to repeated-measures ANOVA.

After the null hypothesis is rejected using the Kruskal-Wallis or Friedman test, *post-hoc* analysis should be performed to determine which comparisons are statistically significant, in the same way that is required after an ANOVA test in normally distributed samples. The Mann-Whitney and Wilcoxon rank-sum tests, respectively, are often performed for these comparisons, although specific multiple comparison tests have been developed for the Friedman ANOVA [9]. As with parametric multiple comparison tests, there is an increased risk of type-I error, and the Bonferroni correction, or a similar approach, should also be applied to control that risk, as demonstrated previously. For example, in Table 4.2 we show the data for

Table 4.1 Indications for parametric tests and each nonparametric analog for continuous variables

Indication	Parametric test	Nonparametric test
Two independent groups	Unpaired *t*-test	Mann-Whitney *U*-test
Two related (dependent) groups' data	Paired *t*-test	Wilcoxon signed-rank test
Three or more independent groups	One-way-analysis of variance (ANOVA)	Kruskal-Wallis rank ANOVA test
Three or more related (dependent) groups' data	Repeated-measures ANOVA	Friedman two-way ANOVA by Rank test

Table 4.2 Clinicopathological and operative data for 819 patients who underwent pancreaticoduodenectomy at Memorial Sloan-Kettering Cancer Center in the 9 years from 2000 to 2008 (Copyrights)

Clinicopathological and operative variables				
		Complications		
Variable	All patients (n = 819)	Yes (*n* = 405, 49.5%)	No (*n* = 414, 50.5%)	*P*-value
Age, years, median (IQR)	67 (58–75)	67 (59–76)	68 (57–75)	0.638
Sex, *n* (%)				<0.001
Female	401 (49.0%)	171 (42.2%)	230 (55.5%)	
Male	418 (51.0%)	234 (57.8%)	184 (44.5%)	
Body mass index, kg/m^2, median (IQR)	26.4 (23.4–29.7)	27.0 (23.8–30.5)	25.8 (22.8–29.2)	0.002
Hypertension, *n* (%)	392 (47.9%)	199 (49.1%)	193 (46.6%)	0.485
Diabetes mellitus, *n* (%)	157 (19.2%)	76 (18.8%)	81 (19.6%)	0.790
Cardiac disease, *n* (%)	189 (23.1%)	107 (26.4%)	82 (19.8%)	0.025
Pulmonary disease, *n* (%)	80 (9.8%)	47 (11.6%)	33 (8.0%)	0.099
Other comorbidities, *n* (%)	533 (65.1%)	258 (63.7%)	275 (66.4%)	0.421
ASA physical status, *n* (%)				0.106
Class 1	17 (2.1%)	6 (1.5%)	11 (2.7%)	
Class 2	422 (51.5%)	201 (49.6%)	221 (53.4%)	
Class 3	370 (45.2%)	190 (46.9%)	181 (43.7%)	
Class 4	10 (1.2%)	8 (2.0%)	2 (0.5%)	
Greatest diameter of tumor, cm, median (IQR)	782 (95.5%)[a]	2.8 (2.0–3.6)	3.0 (2.0–3.8)	0.140
Malignant tumors[b], *n* (%)	701 (85.6%)	342 (48.8%)	359 (51.2%)	0.319
Diagnosis, *n* (%)				
Ampulla of Vater				
Adenocarcinoma	139 (17.0%)	74 (18.3%)	65 (15.7%)	
Adenoma	8 (1.0%)	3 (0.7%)	5 (1.2%)	
Neuroendocrine tumor	3 (0.4%)	1 (0.2%)	2 (0.5%)	
Others	4 (0.5%)	2 (0.5%)	2 (0.5%)	
Bile duct				
Adenocarcinoma	55 (6.7%)	33 (8.1%)	22 (5.3%)	
Neuroendocrine tumor	1 (0.1%)	1 (0.2%)	0	
Others	7 (0.9%)	5 (1.2%)	2 (0.5%)	
Duodenum				
Adenocarcinoma	55 (6.7%)	28 (6.9%)	27 (6.5%)	
Adenoma	9 (1.1%)	5 (1.2%)	4 (1.0%)	
GIST	6 (0.7%)	5 (1.2%)	1 (0.2%)	
Neuroendocrine tumor	4 (0.5%)	2 (0.5%)	2 (0.5%)	
Others	1 (0.1%)	1 (0.2%)	0	
Pancreas				
Adenocarcinoma	437 (53.4%)	197 (48.6%)	240 (58.0%)	
Cystadenoma	23 (2.8%)	14 (3.5%)	9 (2.2%)	
Neuroendocrine tumor	36 (4.4%)	22 (5.4%)	14 (3.4%)	
Pancreatitis	19 (2.3%)	7 (1.7%)	12 (2.9%)	
Others	12 (1.5%)	5 (1.2%)	7 (1.7%)	
Procedure, *n* (%)				
Standard PD	689 (84.1%)	343 (49.8%)	346 (50.2%)	0.702
Pylorus-preserving PD	130 (15.9%)	62 (47.7%)	68 (52.3%)	

Table 4.2 (continued)

Clinicopathological and operative variables

Variable	All patients (n = 819)	Complications		*P*-value
		Yes ($n = 405$, 49.5%)	No ($n = 414$, 50.5%)	
Duration of surgery, min, median (IQR)	266 (221–322)	276 (226–330)	261 (217–313)	0.005
Estimated blood loss, mL, median (IQR)	600 (400–900)	600 (400–1000)	500 (350–800)	0.001
Length of stay, days, median (IQR)	9 (8–13)	12 (9–18)	9 (7–10)	
Any positive margins, *n* (%)	135 (16.5%)	58 (14.3%)	77 (18.6%)	0.109

[a] The total number of measureable tumors was 782

[b] Malignant tumors were confirmed by pathology in 701 patients

ASA American Society of Anesthesiologists, *GIST* gastrointestinal stromal tumor, *IQR* interquartile range, *PD* pancreaticoduodenectomy

clinicopathological and operative variables according to presence of complications in 819 patients who underwent pancreaticoduodenectomy [10]. Nonparametric tests were applied, while continuous variables were presented as medians and interquartile ranges.

Regarding paired analysis, i.e., case-matched, or before-and-after treatment variations for individuals, not by groups, the Wilcoxon signed-rank test is indicated. This test works with cluster data; the concept of clustering is approached in a separate chapter (case-control analysis). Table 4.3 shows an example of case-matched data for postoperative morbidity in 29 patients with initially locally unresectable or borderline pancreatic cancer who underwent resection, compared with data for 29 patients with initially resectable cancer who also underwent resection [11]. Each individual in the unresectable cancer group was compared with their matched control in the resectable cancer group.

4.2.3 Tests of Association for Categorical Variables

Categorical variables are described as proportions or frequencies. Differences in rates between or among categorical variables can be presented as a clear association or not, and can be demonstrated by tests. Differences between groups, as, for example, differences in rates of responders in two different chemotherapy regimens, are usual in oncology for both retrospective and prospective trials. If the proportions or frequencies represent subjects—for instance, deceased or alive patients—instead of ranks or ordinal values, there are specific statistical methods to test those measures of association, with specific indications for when to use the test. Statistical tests of binary variables are used to determine whether a difference in frequencies or proportions between two groups happened by chance. The rules of null hypothesis testing apply here as well.

Table 4.3 Clinicopathological and operative data from 29 patients who underwent neoadjuvant chemoradiation followed by pancreatectomy compared with 29 patients who underwent surgery upfront for pancreas cancer, adjusted by case-matching

Clinicopathological and operative parameters				
	Total	Chemoradiation	Surgery	
Characteristics	$N = 58$ (%)	$N = 29$ (%)	$N = 29$ (%)	P-value
Age (years)[a]	66 (60–72)	64 (61–72)	67 (60–70)	0.77
Gender (Male)	28 (48)	14 (48)	14 (48.)	1
Body mass index[a]	25 (23–28)	25 (22–28)	26 (23–28)	0.58
ASA (higher than 2)	27 (47)	14 (48)	13 (45)	1
Weight loss (%)	5 (0–10)	0 (0–8)	6 (0–10)	***0.003***
CA 19-9 at diagnosis (ng/dL)[a]	170 (49–679)	249 (80–1217)	87 (25–464)	0.33
Tumor site				1
Head	52 (90)	26 (90)	26 (90)	
Tail	6 (10)	6 (10)	6 (10)	
Previous cardiovascular disease	12 (21)	6 (21)	6 (21)	1
Diabetes	6 (10)	3 (10)	3 (10)	1
Pulmonary disease	2 (3.5)	1 (3.5)	1 (3.5)	1
Alb (mg/dL)[a]	4.1 (3.9–4.2)	4.1 (3.8–4.2)	4.1 (4–4.2)	0.49
Hemoglobin (mg/dL)[a]	12.4 (11.3–13.6)	11.7 (10.8–13)	13.4 (11.8–14.3)	**0.03**
Previous surgery	15 (26)	14 (48)	1 (3.5)	***<0.001***
Procedure				1
Distal pancreatectomy	4 (7)	2 (7)	2 (7)	
Pancreaticoduodenectomy	54 (93)	27 (93)	27 (93)	
Vascular resection	8 (14)	4 (14)	4 (14)	1
Operative time (min)[a]	291 (255–335)	271 (251–360)	297 (260–330)	0.88
Estimated blood loss (mL)[a]	600 (350–1000)	600 (300–1000)	600 (355–1000)	0.24
Transfusion	17 (29)	9 (31)	8 (28)	1
Any positive margin	8 (14)	1 (3.5)	7 (24)	0.07
Tumor size (cm)[a]	3 (2.1–3.9)	2.5 (1.5–3)	3.2 (2.8–4.2)	***0.011***
T Stage				***0.016***
(0, 1,2)	7 (12)	7 (24)	0	
3	51 (88)	22 (76)	29 (100)	
N Stage				***<0.001***
N0	31 (53)	27 (93)	4 (14)	
N1	27 (47)	2 (7)	25 (86)	
Any complications	28 (48)	12 (41)	16 (55)	0.42
Any grade 3–5 complications	12 (21)	6 (21)	6 (21)	1
Presence of leaks and fistulae	5 (9)	2 (7)	3 (10)	1
90-Day mortality	1 (1.7)	0	1 (3.5)	1

[a] Median (interquartile). Univariate analyses: McNemar's chi-square and Wilcoxon's signed-rank tests comparing homogenicity between the case and control groups

When the comparison is made with independent samples, the most appropriate test is the chi-square test, which is a nonparametric statistical test. It is the most powerful test for independent categorical samples. There are several uses of chi-square tests in addition to comparisons of frequencies between two groups. It is also possible to use chi-square to verify the comparison between an observed frequency, or proportion, and the theoretically expected parameter for that group. For instance, if you want to determine whether a randomization process was done without bias in a two-group clinical trial, the expected number of subjects in each

Table 4.4 Example of a contingency table

	Group 1	Group 2
With characteristic	A	C
Without characteristic	B	D
Total	n1 (A + B)	n2 (C + D)

group represents 50% of the sample. The chi-square test can then be used to statistically compare the proportion of the observed values and the 50% values expected by chance.

The chi-square test is very easy to perform, as only a contingency table with the number of subjects, as shown in Table 4.4, is required. As the chi-square formula uses all cells—A, B, C, or D in the contingency table—it requires large samples. In some particular situations, the chi-square test might be not appropriate: when sample sizes are <20 [12], when some cell in the table is zero, or when more than 20% of the cells have values <5 [13]. These parameters are not rare but, fortunately, some alternatives are available to face these issues. The first strategy is to collapse some cells to increase the number of observations in each cell. Some groups can be merged into a new 2 × 2 contingency table. However, this strategy has to be based strongly on the study question and must make sense scientifically. It is also important to point out that potential merging groups have to be assigned before data analysis and that it is not recommended to change the statistics after knowing the final results. When the number of subjects in a cell is low in a 2 × 2 contingency table, a statistical correction called Yates's correction for continuity might be used. This mathematical procedure is a conservative strategy used to calculate the chi-square with low values, but it has an increased risk of a type-II error—i.e., it reduces the power of detecting differences between groups. When $n \leq 5$, one can use another test, Fisher's exact test.

Fisher's exact test is a nonparametric test that is often used in comparisons of more than two groups and in situations in which the chi-square test cannot be used, particularly in small samples [12]. A contingency table is required, but the calculation is different from that required for the chi-square test. In the cohort cited previously with 819 patients who underwent pancreaticoduodenectomies (shown in Table 4.2), Fisher's exact test was used for both binomial (gender) and ordinal variables (American Society of Anesthesiology [ASA] score). When the contingency table is bigger than 2 × 2, a statistically significant difference in the comparisons does not demonstrate which proportion is different. Similarly to the one-way ANOVA test for continuous variables, multiple comparisons have to be performed to identify which individual comparisons are statistically significant. But typically for Fisher's exact test for a table larger than 2 × 2, only the global null hypothesis is of interest, not individual comparisons.

For paired comparisons, the chi-square test and Fisher's exact test are not appropriate. In paired analyses it is necessary to take into account the degree of variance of each individual, not the behavior of the entire group. For example, for the comparison of the proportions of cancer patients not indicated for surgery before and after a chemotherapy strategy, these tests are not valid, as values in the two groups are related—patients are the same in the pre- and post-chemotherapy groups. For

Table 4.5 2 × 2 Table for McNemar's test

		Before	
		With characteristic	Without characteristic
After	With characteristic	A	B
	Without characteristic	C	D

these cases, McNemar's test is recommended. A 2 × 2 table has to be performed as well, but the subjects in this table will be represented twice and the table has to be built as demonstrated in Table 4.5. An example of performing McNemar's test in a case-matched analysis is demonstrated in Table 4.3, where binomial variables such as nodal status (N0 versus N1) or the presence of any complications are shown. McNemar's test was used to compare binominal variables (i.e., neoadjuvant chemoradiation therapy versus surgery for pancreas cancer) between the two groups in paired individuals,.

In Table 4.5, A and D represent the subjects whose characteristics did not change during the study and B and C are the subjects whose characteristics changed during the study. For this analysis, B and C subjects are the values of interest. The null hypothesis is rejected using McNemar's test if B and C are sufficiently different.

Conclusion

Suitable analyses of data and the choice of correct tests of measures of association to be used are vital to the research question and may lead to better use of resources and more accurate answers. Although any data can be transformed into categorical variables, tests for continuous variables are more precise than tests for categorical values. Therefore, if the data can be described as a continuous variable, transforming it into a categorical variable can lead to a loss of information and, consequently, a loss of power of the tests [14]. When comparing continuous variables, knowing the type of distribution of the variable is crucial for determining whether to use parametric tests, which are more powerful, or nonparametric tests, which are not influenced by outliers. After tests are performed, p-values have to be interpreted cautiously, because false-positive results can occur, particularly in multiple comparison tests when the Bonferroni correction is not performed; additionally, investigators have to be attentive to the possibility of false-negative results that may arise when an underpowered test (or sample) is used. Another important issue is not to give too much weight to the p-value when it rejects the null hypothesis, as a statistically significant difference may be not clinically significant or may not indicate a therapeutic difference.

In conclusion, for the knowledge of how to handle data and statistics, it might be helpful for researchers to discuss methodology with statisticians, and such discussions are strongly recommended for all research-training programs, so that optimal data analyses and interpretation can be achieved.

References

1. Whitley E, Ball J. Statistics review 3: hypothesis testing and P values. Crit Care. 2002;6:222–5.
2. Baker M. Statistician issue warning on p-values. Nature. 2016;531:151.
3. Razali N, Wah YB. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests. J Stat Model Anal. 2011;2(1):21–33.
4. Kao LS, Green CE. Analysis of variance: is there a difference in means and what does it mean? J Surg Res. 2008;144(1):158–70.
5. Paolino B, et al. Myocardial inactivation of thyroid hormones in patients with aortic stenosis. Thyroid. 2017;27:738.
6. Bewick V, Cheek L, Ball J. Statistics review 9: one-way analysis of variance. Crit Care. 2004;8:130–6.
7. Winer BJ, Michels KM, Brown DR. Statistical principles in experimental design. 3rd ed. New York, NY: McGraw-Hill; 1991.
8. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32(200):675–701.
9. Siegel S, Castellan NJ. Nonparametric statistics. 2nd ed. Belmont, CA: Wadsworth Publishing; 1989.
10. Araujo RL, Karkar AM, Allen PJ. Timing of elective surgery as a perioperative outcome variable: analysis of pancreaticoduodenectomy. HPB (Oxford). 2014;16(3):250–62.
11. Araujo RL, Gaujoux S, Huguet F. Does pre-operative chemoradiation for initially unresectable or borderline resectable pancreatic adenocarcinoma increase post-operative morbidity? A case-matched analysis. HPB (Oxford). 2013;15(8):574–80.
12. Winters R, Winters A, Amedee RG. Statistics: a brief overview. Ochsner J. 2010;10:213–6.
13. Cochran WG. Some methods for strengthening the common x^2 tests. Biometrics. 1954;10:417–51.
14. Portney LG, Watkins MP. Foundations of clinical research. 3rd ed. Philadelphia, PA: F.A. Davis; 2015.

5 Sample Size Calculation in Oncology Studies

Rachel P. Riechelmann, Raphael L. C. Araújo, and Benjamin Haaland

5.1 Introduction

Clinical trials of new cancer-directed therapies, including oncological surgery, set the basis for the progress of cancer research. They represent the tools that determine the standard of care. Therefore the designs of clinical trials have to be thoroughly planned to deliver the most accurate and reliable results. There should be careful planning of the trial design, number of arms, eligibility criteria, the primary and secondary endpoints, statistical analyses, and sample size. At the very core of the study design is sample size calculation (SSC).

SSC is defined as the computation of the minimum number of participants to be included in a study in order to detect a true effect or value. In most cases this effect is a pre-determined estimated difference between related or unrelated groups. The final sample is supposed to be representative of the general population. Such calculation must be performed *a priori* and it takes into account the chances of false-positive or false-negative study results, among other factors. Hence, several assumptions must be made for SSC that directly influence the study results.

This chapter will summarize general concepts of SSC for clinical trials and common designs in non-intervention studies in oncology. It is not our aim to go over the mathematical formulas behind SSC, but rather to provide the basic concepts for clinical investigators to compute the sample sizes of their own studies.

R. P. Riechelmann, M.D., Ph.D. (✉)
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil

R. L. C. Araújo, M.D., Ph.D.
Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos Cancer Hospital, Barretos, SP, Brazil

B. Haaland, Ph.D.
University of Utah, Salt Lake City, UT, USA

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_5

5.2 Fundamental Steps for Sample Size Calculation in Comparative Studies

The SSC depends on the type of variable used for the primary endpoint (continuous, binary, time-to-event), the type of analyses (intergroup or intragroup), the variability of the primary endpoint (measures of dispersion), the magnitude of difference between the study groups (delta), the type of study (superiority randomized clinical trial, cohort study), the target probability of false-positive and false-negative results, and the expected loss of data (dropout rate). Each parameter for SSC is discussed in detail below, and Table 5.1 provides a summary of the components demanded for SSC, as well as summarizing how they influence the final sample size.

5.2.1 Defining the Primary Endpoint

The rationale for supporting the SSC of a study starts with the formulation of the research question. The next step is to define the study primary endpoint, i.e., the measure of efficacy. Notably, the study objective is not synonymous with the study

Table 5.1 Required parameters for sample size calculation (SSC) in comparative studies and their influence on the final sample size [4]

Required parameters for SSC	Characteristics	The influence of parameters on the sample size
Type of primary endpoint	Categorical, numerical, time-to-event	Influences the sample size in the context of other parameters
Study design	Single-arm, randomized, etc.	Randomized trials tend to have larger samples
Test of hypothesis	Superiority, non-inferiority, or equivalence	Non-inferiority/equivalence trials have larger samples
Expected outcome of standard therapy	Set the basis for null hypothesis	Alone does not influence the sample size
Expected outcome of experimental therapy	Determination of delta; non-inferiority margin	Alone does not influence the sample size
Type-I (alpha) and type-II (beta) errors	Probabilities of false-positive and false-negative findings	The larger the errors, the smaller the sample
Power	1—beta	The higher the power, the larger the sample
Observed statistical significance	*P* value	The smaller the *P* value, the larger the sample
Direction of statistical test	One-tailed or two-tailed	Two-tailed studies have larger samples, considering the same significance level
Dropout rate	Often 10%	To be added after SSC
Length of follow-up and accrual	For time-to-event variables only	The longer the periods, the smaller the sample
Measure of dispersion	For numerical variables only; reflects variability	The larger the dispersion, the larger the sample

endpoint. The objective conveys *what* we want to achieve, while the endpoint determines *how* we aim to achieve the objective. For instance, in a study whose objective is to evaluate drug efficacy, the primary endpoint can be progression-free survival from the date of randomization or the disease control rate at 8 weeks from treatment initiation, as measures of efficacy. While a study may have more than one primary endpoint, only one endpoint is used for SSC. Frequently used primary endpoints in oncology are the response rate (RR), progression-free survival (PFS), disease-free survival (DFS), overall survival (OS), and patient-reported outcomes (PROs). The RR is defined as the proportion of patients whose tumors decrease by a fixed benchmark percentage compared with a baseline measurement. PFS is defined as the time between the first day of treatment administration or date of randomization until the date of disease progression or death, whichever comes first; patients who are lost during follow-up or those who did not experience the event at the time of analyses are censored. DFS is similar to PFS, but is used in the adjuvant setting. OS is counted from the date of randomization or of treatment initiation until death from any cause. PROs relate to endpoints notified by patients, such as pain, fatigue, nausea/vomiting, and quality of life—these are mostly measured by validated questionnaires.

The type of variable used for the primary endpoint directly influences the SSC and consequently, the final sample. Variables are generally classified as quantitative or qualitative (see Chap. 3). Qualitative variables can be categorical or ordinal. Categorical (also known as nominal) variables are often binary or dichotomous, meaning that they are "yes or no" according to a given parameter. The following are examples of binary variables: the RAS mutation status of a tumor is either mutated or wild type; to determine the response rate, each patient is classified as responder versus non-responder; in DFS at 3 years, patients are either with or without disease at this time point; the disease control rate reflects the proportion of patients with or without tumor progression at a given time point. In qualitative ordinal variables an increasing or decreasing order exists. Examples of ordinal variables are the Eastern Cooperative Oncology Group (ECOG) performance status scale (from 1 to 5) and the Common Toxicity Criteria for Adverse Events grading of toxicity (from 1 to 5). Quantitative variables are numerical and can be discrete, such as counts, or continuous, where a figure can assume infinite possibilities in a given interval. Examples of discrete quantitative variables are number of children, number of metastatic sites, days of hospitalization, quality-of-life scores, and the ki67 labeling index. Examples of continuous quantitative variables are blood pressure, hemoglobin level, and body weight. Numerical variables are associated with imprecision, i.e., variability is implied in their result. Such variability is often called dispersion, and can be conceptualized as the interval or spread where the numerical variable is encountered. Common measures of dispersion for numerical variables are variance, standard deviation, and interquartile range.

A third type of variable is the time-to-event outcome. These variables are unique, as they do not encompass complete data, because, for the outcome to be recorded, a sufficient length of follow-up is necessary. Thus, survival time will be unknown for many patients in a study if the follow-up is not long enough. Such patients are censored, i.e., they have not experienced the main outcomes (progression, death, recurrence) during the study period or were lost to follow-up for different reasons. Censored survival times may underestimate the real time to the event of interest if the censored data is not handled properly [1]. Unlike the SSC for binary and continuous variables, which provides the required number of patients, the SSC for time-to-event variables delivers the required number of *events*, not patients. To compute the number of patients, we need to consider the event rate being studied. Knowing the biological course of the disease, investigators can estimate the follow-up and accrual periods that allow events to occur. That is why the lengths of follow-up and accrual also need to be considered for SSC in studies where the primary endpoint is a time-to-event variable. Both these dimensions of time directly influence the SSC, because if the follow-up and accrual periods are short, patients might not have time to achieve disease-related events, and then they are censored. In such cases, the sample has to be larger to take into account the shorter time available for events to occur. For example, trials of indolent tumors need sufficient follow-up to observe disease progression. Longer follow-up and enrollment periods allow smaller samples because there is enough time for events to arise. For example, trials of treatments for inoperable multiform glioblastoma may have shorter follow-up times because disease-related events occur in the short-term, due to the poor prognosis associated with this cancer.

Some endpoints can be measured either as a continuous (or time-to-event) variable or as a dichotomous variable. For example, a study in ovarian cancer wants to look at biochemical response, defined by reduction in CA 125 serum levels. The CA 125 response can be defined as either the mean decrease relative to baseline (endpoint is a continuous variable) or as the proportion of patients who showed a decrease of at least 30% of baseline CA 125 levels (endpoint is a categorical variable). The required sample size is larger when the primary outcome is binary (responder versus nonresponder) as compared with continuous or time-to-event endpoints, because information is lost when continuous data is bucketed into two categories [2].

5.2.2 Tests of Hypotheses

After determining the primary endpoint—and type of variable—researchers have to consider the expected outcomes, in the control and experimental groups, which reflect the magnitude of the difference between the study arms. Of note, having a control and an experimental group does not necessarily imply that the SSC is for a randomized controlled trial. Indeed, these two groups are determined for SSC assumptions only in order to estimate the magnitude of difference with a new intervention, and so these rules apply to uncontrolled trials and retrospective studies. Such assumptions set the basis for hypothesis testing.

A hypothesis test is a test performed in an experiment that will show if there is strong evidence in favor of a claim. All tests of hypotheses involve making an initial assumption, collecting the data, and then deciding whether to fail to reject or reject the initial assumption. For example, we hypothesize that R0 resection of lung and/or liver metastases from colorectal cancer improves overall survival when compared with systemic chemotherapy in a molecularly unselected group of patients. This implies the expected outcomes in the control and experimental groups. Based on the survival results of the resected group, the investigators will either reject or accept their initial assumption that resection improves survival in metastatic colorectal cancer. If they find a 2-month difference in median survival between the groups, they may accept their survival assumption. However, this finding may not necessarily be the truth; it could have happened only by chance. On the other hand, if researchers reject their survival assumption about metastasectomy, meaning the intervention does not improve survival, this may be a false-negative finding by chance. Therefore, for any hypothesis there are probabilities of false-positive and false-negative results. These are called type I or alpha errors and type II or beta errors, respectively. The hypothesis that investigators base their assumption on is called the null hypothesis or H0 (in this example it means that resection and chemotherapy offer similar survival times), while the alternative hypothesis or H1 reflects the opposite of the null hypothesis (resection offers a survival time different from that with chemotherapy) (Fig. 5.1). The anticipated magnitude of the difference between H0 and H1 is called delta.

Let's have another example. We postulate that drug A (experimental) improves the RR when compared with drug B (control) in patients with metastatic synovial sarcoma. Given that the known RR offered for drug B is approximately 15%, we assume drug A will provide an RR of 30%. The null hypothesis is that drug A is not better in terms of RR than drug B; the alternative hypothesis is that drug A provides a higher RR than drug B. So H0: drug A = drug B, both with an RR of approximately 15%, and H1: drug A > drug B. Here the delta or effect size is 15%. However, because it is impossible to prove that two variables are mathematically equal, given that the study sample does not represent the whole population, a statistical hypothesis test assesses the evidence *against* the quantities of interest being equal. It has been a consensus in the scientific community that it would be dangerous to consider and standardize a falsely positive treatment because it could harm patients. In contrast, it is not considered to be "hazardous" to leave out a good therapy that was falsely considered ineffective. Conventionally, most tests of hypothesis use a 5% value for the alpha error (type I error rate) and 10–20% for the beta error (type II error rate). Because the statistical power is defined as 1 minus beta (1 – β), most studies set the power as 80–90%. Therefore, the power of

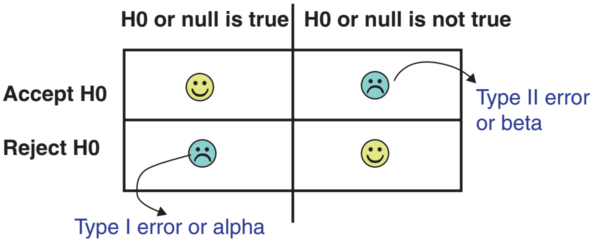


Fig. 5.1 Schema of the relationship between reality (H0 true or false) and the outcome of a hypothesis test (reject or fail to reject H0)

a study indicates the probability of rejecting the null hypothesis when H1 is true. In other words, the power provides the probability of detecting a real difference of a particular size between two groups when this difference exists. For example, a single-arm phase II trial demonstrates a median PFS of 12 months for a new therapy; the median PFS reported by historical data with another treatment is 8 months. This difference is statistically significant and the power was set at 85%. This means that the statistical hypothesis test was constructed so that if a difference of this magnitude were actually present, then the hypothesis test would have an 85% chance of discovering it.

The delta, also called effect size, represents the estimated magnitude of difference between the groups. The delta is a critical part of the SSC, because flawed assumptions can lead to false results. Intuitively, small deltas lead to large samples, while large deltas result in smaller samples. Therefore, large trials are needed to detect small differences, but small differences, on the other hand, may not be clinically meaningful. For example, a phase III trial with 4000 patients with metastatic adenocarcinoma of the pancreas shows a survival benefit of 15 days, with *P* value = 0.00001; although the result is statistically significant, this is clearly not clinically relevant because the median OS with the standard arm in this setting was in the range of 6–7 months. In contrast, small trials may be underpowered to detect clinically and statistically significant differences if the delta assumption is too ambitious for the clinical setting. For example, a randomized controlled phase II trial is designed to find an absolute gain of 40% in RR, but the results turn out to be negative, with a difference of 25%. Forty percent absolute benefit is certainly a large estimated delta. While the result was not statistically significant, because the trial was underpowered to identify absolute differences smaller than 40%, a 25% reduction in RR might still be clinically important.

The determination of delta relies on information about the outcomes of the control and experimental groups. Assumptions about the control group should come from data reported by studies with sound methodology, such as randomized trials and/or meta-analyses of controlled trials. Investigators must also be attentive about whether the population of their study resembles the population from the published literature. This is because differences in patients' prognosis and tolerability may lead to varying results in different settings. For example, if an SSC is planned taking into account the expected survival of a control treatment reported in trials that included only patients with ECOG 0 or 1, the assumptions for the control arm might be flawed, and likely overestimated, if investigators allow ECOG 2 patients to be enrolled. Or if the published literature about a standard-of-care therapy comes mostly from Asian countries, one cannot infer that the treatment-related toxicity profile will be the same in Latin America. In certain situations, there is very little solid evidence relating to a treatment. In these cases, institutional databases can be used to estimate the outcomes of the control group. If such databases are also unavailable, clinical judgment is advised; discussion with experts, extrapolation from data from similar interventions, or even personal experience can be employed in these instances, although this has to be clearly stated in the protocol/manuscript. Likewise, estimations about the outcomes of the experimental intervention have to be carefully planned based on the scientific evidence from other studies that tested that intervention, studies that evaluated similar therapies and disease settings, realistic assumptions of potential benefit, and clinical judgment (it is also called the minimum clinically meaningful difference).

The delta can be estimated as either an absolute or a relative difference. For trials where the RR, disease control rate, or DFS rate at 5 years are the primary endpoints, absolute differences can be considered so as to estimate the magnitude of difference. In trials that use survival-type endpoints, the delta is determined as a relative difference, often as the hazard ratio, which measures the relative risk or hazard (of death or progression, for example) throughout the study.

The *observed* significance level is also an important component of the SSC. It is generally called the *P* value, and reflects the probability that the observed difference (or a more extreme difference) in the outcomes of two groups (or observed delta) is due to chance, if in fact the null hypothesis that there was not any difference between the groups is true. The *P* value summarizes how consistent the observed data is with the null hypothesis. A small *P* value implies that the data is not consistent with the null hypothesis, and in turn provides evidence that the null hypothesis is false. For instance, a case-control study comparing two groups observes a difference in proportions of patients with poor ECOG status, and a corresponding *P* value of 0.03 (as compared with the typical statistically significant benchmark of <0.05). In this case, there is a 3% probability of observing data this extreme or more extreme under the assumption that the proportions of poor ECOG status are the same in the two groups. While the magical *P* value <0.05 has been widely accepted by the scientific community as the threshold for determining statistical significance, researchers are free to use other *P* value cutoffs for statistical tests. In fact, if multiple tests are undertaken, the *P* values need to be adjusted for multiple comparisons to avoid false-positive findings.

To compute a sample size researchers have to determine whether the study is one- or two-tailed (or -sided) based on statistical significance or the *P* value (Fig. 5.2). For a one-sided *P* value, the test has one direction, i.e., it aims to determine whether H1 > H0—i.e., drug A offers a higher RR than drug B, for instance. For a two-sided *P* value, the test has two directions, i.e., H1 > or < H0 or H1 ≠ H0—i.e., the RR provided by drug A is different from the RR offered by drug

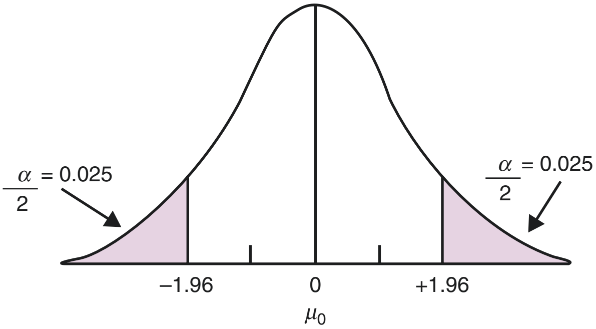


Fig. 5.2 One- or two-tailed statistical significance

B, being either higher or lower. Two-tailed tests are the rule in clinical cancer research because they provide information on whether the experimental treatment is better or worse, giving more reliability to trial results. One-tailed studies can be used in highly specific scenarios of superiority oncology trials, where investigators are confident that the intervention can only be better than or similar to the standard therapy, but that it is very unlikely to be inferior. This has to be convincing, based on solid scientific evidence. Other studies where one-sided tests are justified include studies designed to evaluate safety or toxicity, risk evaluation, and laboratory research [3]. Non-inferiority trials are always one-tailed, as discussed further.

Confidence intervals, usually set at 95%, are another way of assessing statistical significance. They indicate a plausible range of values for the true difference in the population, regardless of the type of endpoint variable. If a confidence interval for a mean difference does not contain zero, then the corresponding hypothesis test that the mean difference equals zero is rejected. On the other hand, if the confidence interval contains zero, then the test that the mean difference equals zero cannot be rejected.

All statistical inferences rely on hypotheses tests of superiority, non-inferiority, or equivalence, as discussed in the following sections.

5.2.3 Dropouts

In prospective studies, it is common that the number of subjects with analyzable data at the end of the study is less than the total number recruited at the beginning. This is because patients drop out during the study period for various reasons, such as moving to another city, withdrawing consent, and death. The dropout rate varies according to the type of study and patient population. For example, clinical trials of refractory cancer patients receiving supportive care exclusively may have a high attrition rate because these patients are very sick and may not be well enough to come for assessment. Prospective cohort studies, especially those with long follow-up times, may also lead to high losses of study patients. Examples of such studies include cancer survivorship cohorts, vaccine trials, and cancer prevention studies. The phenomenon of dropout comprises a very important practical aspect of SSC planning because it directly influences the power of the study. If less than the anticipated number of patients experience events, the study may be underpowered to measure differences in outcomes.

The average dropout rate in clinical oncology trials is 10%. For scenarios with a poor prognosis, dropout rates of 20% or even 30% should be expected. Having said this, after the SSC has determined the number of patients to be enrolled, investigators should add an “overhead” of 10–30% on top of that number, depending on the study type and setting (Table 5.1).

5.2.4 Study Designs and Types of Comparisons

The study design directly influences the sample size. For example, observational studies often need hundreds of patients to provide an acceptably narrow confidence interval around the true value. Studies that evaluate the pre- and post- effect of an intervention in a single group of patients usually require half the sample size that would be required if the study had an independent control group. A clinical trial with a two-tailed hypothesis demands more patients than a one-tailed study. Non-inferiority clinical trials require especially large samples because the delta is very small. The prognosis of the study population also influences the SSC calculation because patients with a higher risk of experiencing the event (death, progression) will experience the event sooner. In such cases the study period (accrual and follow-up) is short, the number of events is high, and consequently, the sample size is smaller in comparison with the sample size in studies of indolent tumors. In the following section we discuss the particularities of the most common study designs utilized in oncology.

5.3 Sample Size Calculation in Oncology Clinical Trials

5.3.1 Single-Arm Phase II Trials

The SSC for single-arm oncology trials can be basically performed following two classical designs: the Fleming design [5] and the two-stage design and its variations. In the Fleming design, the hypothesis test determines the magnitude of difference between the expected outcomes (often RR) associated with the experimental and with the standard therapies. The determination of the outcome of the standard intervention is based on historical data, i.e., information from publications or institutional retrospective databases. The required parameters for SSC for single-arm phase II trials are the delta, type I and II errors (and correspondingly power) and the definition of a one- or two-tailed test.

The phase II two-stage design envisioned by Gehan [6] and optimized by Simon [7] was developed to screen out drugs which are not efficacious, enrolling a smaller number of patients than the Fleming design. This design commonly enrolls a first stage of 10–15 patients; if no response or an insufficient number of responses is seen, the probability of success is very low and the trial is terminated at this stage. In contrast, if at least one (sometimes more) response is observed, a second stage of enrollment is carried out until a pre-defined number of patients is achieved. The sample size is based on RR probabilities considered to be “futile” or “promising” for developing further studies. The number of patients per stage is determined based on the null and alternative hypotheses, with associated delta, type I and II errors, and power. The number of patients per stage has been calculated by Simon and reported in tables [7]. The example shown below and in Table 5.2 is an extract from a table depicting the optimized and minimax two-stage designs [7], where $p0$ is the expected RR associated with the standard therapy, $p1$ is the expected RR with the new therapy, the delta ($p1 - p0$) is 20%, ≤r1/≤n1 is the minimum number of responses ($r1$) among the estimated number of patients ($n1$) required to continue to the study second stage, r/n is the upper limit of the number of responders in the second stage, (r) determines whether the study is negative, (n) is the final sample, and EN ($p0$) reflects the average number of patients enrolled. The probability of early termination (PET)

Table 5.2 Design for $p1 - p0 = 0.20$

		Optimal design				Minimax design			
		Reject drug if response rate				Reject drug if response rate			
p_0	p_1	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)
0.05	0.25	0/9	2/24	14.5	0.63	0/13	2/20	16.4	0.51
		0/9	2/17	12.0	0.63	0/12	2/16	13.8	0.54
		0/9	3/30	16.8	0.63	0/15	3/25	20.4	0.46
0.10	0.30	1/12	5/35	19.8	0.65	1/16	4/25	20.4	0.51
		1/10	5/29	15.0	0.74	1/15	5/25	19.5	0.55
		2/18	6/35	22.5	0.71	2/22	6/33	26.2	0.62
0.20	0.40	3/17	10/37	26.0	0.55	3/19	10/36	28.3	0.46
		3/13	12/43	20.6	0.75	4/18	10/33	22.3	0.50
		4/19	15/54	30.4	0.67	5/24	13/45	31.2	0.66

PET probability of early termination. See text for definitions of other terms used in the Table column heads

is when the true response probability is $p0$, and each of the three lines in the Table provides the number of patients per stage according to alpha and beta error simulations – 0.10 and 0.10, 0.05 and 0.20, and 0.05 and 0.10, respectively (Table 5.2).

5.3.2 Multiple-Arm and Randomized Phase II Trials

The above SSC used for single-arm phase II trials also applies to certain types of multi-arm phase II trials, such as biomarker-driven trials. These trials are phase II trials with multiple study arms, where patients are molecularly selected to receive directed therapy. In these trials, the SSC is done for each arm as if for single-arm trials, not making assumptions for intergroup comparisons.

Randomized phase II trials are comparative clinical trials designed to evaluate the preliminary efficacy of anticancer agents. They should not establish a new standard of care because they are not powered to detect differences in true oncology endpoints, such as overall survival or quality of life. Randomized phase II trials are designed to screen out inefficacious drugs in a more precise setting, where randomization controls for systematic errors [8, 9]. Therefore their SSC is not meant to measure differences between the study arms, but rather to estimate the effects within each arm, as if several individual single-arm phase II trials were collapsed into one trial through randomization. Randomized phase II trials that fall into this category are the "pick the winner" design and the randomized controlled trial (see Chap. 11).

Randomized discontinuation trials and randomized phase II/III trials are distinctive in that they are designed to statistically evaluate differences between study groups. In this regard, their SSC resembles those of phase III trials, discussed in the next section, except that they utilize surrogate rather than true endpoints of benefit in oncology, such as the RR and PFS.

Randomized crossover trials are randomized phase II or III trials with two study arms, where patients in both arms receive the study intervention sequentially, each arm crossing over to the other arm. In oncology, crossover trials can be used to investigate treatment sequencing, which is important for the evaluation of mechanisms of drug resistance. This design offers high internal validity because both groups are exposed to both treatments, which allows intragroup comparisons, i.e., pre- and post- evaluations, where each patient is compared with him/herself, with consequently less variability. This is called paired analysis or comparison of dependent groups. Because of their lower variability, crossover designs result in smaller sample sizes than trials that compare independent groups. Sample size software calculators generally ask whether the comparison is between independent or dependent groups.

5.3.3 Sample Size Calculation for Randomized Phase III Trials

Phase III clinical trials are designed to test whether a new experimental intervention is better than, non-inferior to, or equivalent to the old treatment. Hence, these studies test a hypothesis of superiority, non-inferiority, or equivalence. While the SSC assumptions for all types of phase III trials follow the rules for the type of primary endpoint variable, effect size, type I and II errors, and determination of statistical significance, there are some differences according to the type of study.

Superiority phase III trials test whether one intervention is better than another. However, to be more conservative and methodologically sound, superiority trials test whether an intervention is *different* from another, i.e., whether it is better or worse. That is why most phase III superiority trials are two-sided: they test two directions, i.e., H1 ≠ H0. The downside of a two-tailed study is that the sample is larger than that in a one-tailed study. Only in very specific situations may investigators be confident enough that the experimental therapy can only be better, or very unlikely worse, than the old treatment. If a one-tailed phase III trial shows negative results, then the correct interpretation of the one-sided hypothesis is that there is insufficient evidence to conclude that the new treatment is better than the standard. This may lead to some clinicians misinterpreting the result and believing that the new intervention is similar to the old, rather than being inferior. Therefore regulatory agencies require that researchers design two-tailed instead of one-tailed phase III superiority trials. For example, the phase III trial REAL-3 of panitumumab added to the standard first-line EOC chemotherapy regimen (epirubicin, oxaliplatin, and capecitabine) in patients with advanced esophagogastric adenocarcinoma was a superiority study powered to detect a 10% improvement in the overall survival rate at 1 year, from 45% to 55%, with a hazard ratio (HR) of 0.749. The study was designed with a 10% type II error and a two-sided alpha of 0.05. The primary endpoint result for overall survival showed an HR of 1.37, with a 95% confidence interval of 1.07–1.76 for the panitumumab arm [10]. This finding demonstrates that adding an anti-epidermal growth factor receptor (EGFR) agent in this setting was detrimental, leading to a higher risk of death in comparison with the standard EOC regimen. This example highlights the importance of using a two-sided design.

In contrast, non-inferiority phase III trials are always one-tailed. They test only one direction, i.e., whether H1 < H0, by a certain magnitude, which is called the non-inferiority margin (NIm). The NIm is the benchmark to which one compares the lower boundary of a 95% confidence limit on the HR, or risk ratio. The NIm is supposed to be small enough so that the experimental intervention is not considered to be clinically inferior to the control treatment, but just *slightly* inferior. For SSC, the delta must be less than the NIm. The non-inferiority may be acceptable if the new therapy provides relevant benefits that outweigh the marginally inferior result, such as more convenient schedules, lower cost, and/or less toxicity [11].

Selection of the NIm is critical for the SSC of non-inferiority trials. It should take into account clinical judgment, solid knowledge of the effect of the new therapy, and the assumption that the experimental treatment is more efficacious than placebo [11]. Differently from superiority trials, the hypothesis test for non-inferiority trials considers the null hypothesis to be H0 – H1 ≥ NIm and the alternative hypothesis to be H0 – H1 < NIm. If the results of a phase III non-inferiority trial show that its NIm lies within the pre-defined confidence interval, the null hypothesis is rejected and the study is considered statistically significant; on the other hand, if the NIm falls beyond the lower boundary of the pre-specified confidence interval, the null hypothesis fails to be rejected and the study is negative because the new therapy may in fact be inferior to the control treatment. Because non-inferiority trials (should) use small deltas, they often have large samples. Non-inferiority trials are also analyzed differently from superiority phase III studies; the latter should be primarily analyzed on an intention-to-treat principle, while non-inferiority trials are commonly analyzed on both an intention-to-treat and a per-protocol basis (see Chap. 13).

Equivalence phase III experiments are two-sided trials that test whether a new intervention is similar to the control therapy. “Similar” is defined by an equivalence margin (Eqm), which is a range of true effect from being marginally superior to being marginally inferior. The null hypothesis in equivalence trials is that the two groups are not equivalent, while in the alternative hypothesis, the groups are equivalent. So the null hypothesis is H0 – H1 ≥ Eqm or H0 – H1 < Eqm [12]. Similar to the NIm, the determination of the Eqm is critical for SSC and should be carefully planned based on clinical judgment and previous well-designed trials.

Figure 5.3 summarizes the boundaries used to determine statistical positivity in phase III trials.

Factorial designs comprise a different type of phase III superiority trial, where two (or more) interventions are evaluated in the same study by pooling the effect of one strategy against the others. In a 2 × 2 design, for example, study subjects are randomized to receive intervention A, intervention B, both, or none. Owing to the complex planning required, analyses and interpretation of factorial designs have not been widely implemented in oncology. Nonetheless factorial designs confer advantages over standard parallel designs, because they are capable of investigating the effect of two different therapeutic strategies in the same

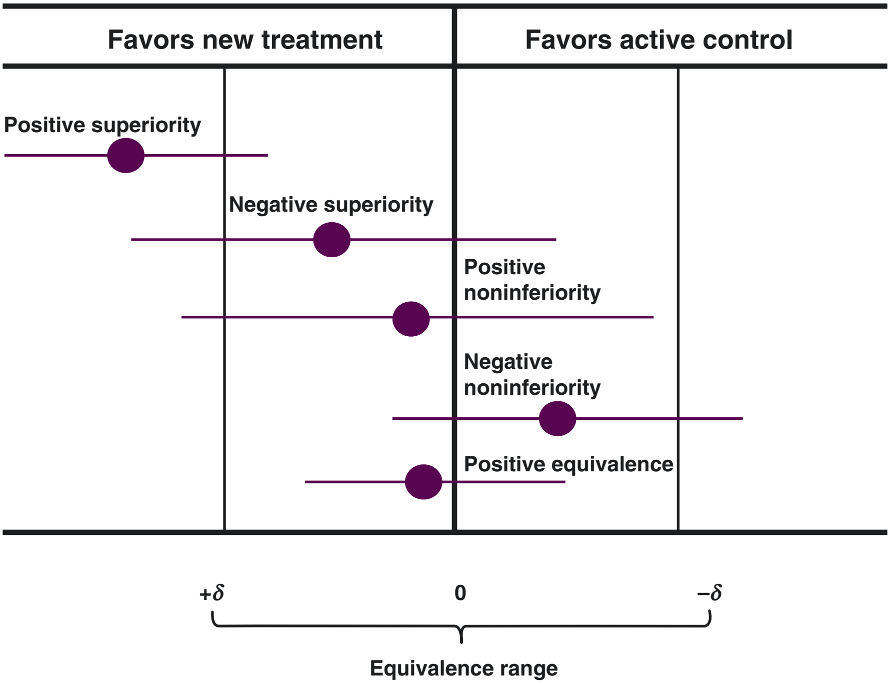


Fig. 5.3 Representation of superiority, non-inferiority, and equivalence trials

study and the effect of each intervention separately or combined. The SSC for factorial designs may be similar to that of standard phase III superiority trials, when independent calculations are made based on effect sizes for each of the interventions compared with their respective controls; then researchers select the larger sample so the trial is considered to be powered to detect the *main effect* of each intervention [13]. However, this sample size is based on the assumption that there are no interactions (or influential effects) between the interventions. If there is interaction, the sample size needs to be adjusted [14]. A classical example of a factorial design is the EORTC 22921 trial [15], where patients with localized rectal cancer were randomly allocated to four independent groups: preoperative radiation, preoperative chemoradiation, and each of these with or without further adjuvant 5-fluorouracil (5FU) chemotherapy. The primary endpoint was the overall survival, which compared the two preoperative groups versus the two postoperative groups. The SSC was performed as if it were a conventional phase III superiority trial, comparing the two mentioned groups for an expected absolute difference of 10% in overall survival rate at 5 years, assuming a power of 80% and a two-tailed alpha of 0.05. In another example, in the REAL-2 phase III trial, patients with metastatic esophagogastric cancer were randomized twice: to receive epirubicin and cisplatin or oxaliplatin (to test for the best platinum agent) and epirubicin plus fluorouracil or capecitabine (to test for the best fluoropyrimidine) [16].

5.3.4 Particularities of Sample Size Calculation for Randomized Phase III Trials

5.3.4.1 Stratification

Phase III cancer clinical trials are balanced, with respect to uncontrolled factors that could impact outcomes, via randomization. Stratification, on the other hand, permits the grouping of patients according to prognostic (age, sex, molecular alterations) and non-prognostic (clinical research center) factors that could influence the trial outcomes. In each stratum subjects are randomly allocated to study arms through pre-defined randomization criteria. Once the trial is completed, each stratum represents a subgroup that is then analyzed. While there is no limit to the number of stratifications, fewer strata lead to more patients in each stratum. The difference between stratified analysis and subgroup analysis is that the latter is done retrospectively, and commonly in an unplanned manner, while for the first, *a priori* characterization of strata and prospective collection of data are carried out. Regardless of timing, retrospective subgroup or stratified analyses are typically regarded as hypothesis-generating, and need to be tested in further research. The results of retrospective analyses are much less trustworthy than those of stratified analysis, owing to inflated type-I error rates, and can be risky to use for treatment decisions.

The importance of performing stratified analyses relies on the attempt to identify any confounding factor or outlier result that could interfere with the overall results; in other words, stratification tries to correct for potential prognostic imbalances. The likelihood of group imbalances decreases with increasing sample sizes, supporting the argument that trials with hundreds of patients do not need to be stratified [17]. Despite its importance for trial design, stratification is rarely contemplated as a parameter for SSC in most oncology trials. While stratified analyses can reduce the sample size of equivalence trials, they do not influence the SSC for superiority or non-inferiority phase III clinical trials [17].

5.3.4.2 Unequal Treatment Allocation

The use of placebo exclusively as the control arm is common in randomized trials in oncology, particularly in patients with metastatic refractory disease. Because of the ethical implications and concerns about poor accrual, some trials allocate more patients to one arm than the other. In general, these trials perform 2:1 or 3:1 randomization, giving patients with advanced cancer more chance to receive the active treatment rather than placebo exclusively. The SSC for unequal allocation follows the same rules as those applied to superiority phase III trials, except for the distribution of the sample in a 2:1 or 3:1 ratio. For example, the RECOURSE trial was a placebo-controlled phase III trial of TAS102, an oral antimetabolite chemotherapeutic agent, in patients with refractory metastatic colorectal cancer [18]. The study randomized 800 patients in a 2:1 proportion. Because of the 2:1 ratio, 800 was divided by 3 ($N = 266$), allocating two-thirds of the patients to the active group ($N = 534$) and one-third to the placebo arm ($N = 266$).

5.3.4.3 Interim Analyses

Clinical trials are usually longitudinal studies that accumulate data over time. With the aim of monitoring for preliminary evidence of efficacy or futility, early stopping rules have been incorporated into most phase III registration trials in oncology. This is done through interim analyses, which allow a group of independent reviewers (Data Safety Monitoring Board) to look into the data and, based on pre-specified rules, decide whether to stop the trial because of indubitable efficacy findings, stop the trial because of futile results, or continue the trial until the next planned analysis [19]. Interim analyses must be planned before starting the trial to ensure research integrity.

There is a great debate on the ethics of interim analyses, because stopping a trial prematurely owing to efficacy results may save further patients from receiving a futile therapy, but at the same time, an interrupted trial may be underpowered to evaluate overall survival gains, for example, the evaluation of which is conditional for regulatory approval in some countries. Without proper adjustment of the stopping boundaries, the more interim analyses performed, the higher is the chance of false-positive results. Hence, several methods have been developed to adjust for alpha spending function. In general, these methods decrease the level of statistical significance in order to control for type-I error, without reducing the study power, thereby making it more difficult to reject the null hypothesis at each interim analysis [19].

The sample size may or not be re-estimated following interim analyses results. In general, special techniques are needed to preserve the type-I and type-II error rates if the sample size is adjusted based on interim results. A common technique with good performance is to use two stages. At the end of the first stage, a one-sided level α_1 (we will discuss choosing α_1 in a moment) test of efficacy is performed. If this test fails to reject, then the sample size is re-estimated for a new level of significance (α_2), which depends on the first stage P value, and is based on the current data, except that the target difference, or delta, remains fixed throughout. The second stage sample size can be taken to achieve the desired power, conditional on the first stage data. Common parameters that might be estimated based on the initial sample include variances, event rates, and dropout rates. The relationships between the first and second stage levels of significance and the overall level of significance are described by simple formulas. For more details see [20].

5.4 Sample Size Calculation for Bayesian Adaptive Designs

Broadly, Bayesian techniques work by using the current data to update prior beliefs, as described by the posterior distribution. Bayesian adaptive designs, in particular, use posterior predictive probabilities, such as the probability that each arm is the best or the probability of arm-specific or overall futility, to inform trial decisions at each stage, such as the allocation of patients to arms, or the decision for arm-specific or overall trial termination. Bayesian adaptive designs are typically characterized by both the need for extensive simulation to determine a trial's operating properties and the sophisticated modeling of dose-response curves and related endpoints. Bayesian adaptive designs have increased in prominence because their seamless incorporation of prior information and sophisticated modeling of patient trajectories ordinarily result in both small sample sizes and more precise identification of the most promising treatment.

5.4.1 Fundamental Concepts for Sample Size Calculation in Non-comparative Studies in Oncology

Non-comparative studies, such as retrospective series, prospective cohorts, and cross-sectional surveys, are very common in oncology. They are designed to evaluate outcomes in a given population, assuming that the study sample is *representative* of a population of interest. While SSC is not mandatory for every non-comparative study, it is advisable for study planning so researchers do not unnecessarily collect extra or collect insufficient data. However, it may be justifiable to leave out SSC in studies of rare diseases or when there are budget/sample constraints. In these cases, convenient and feasible samples are used and in some instances, power calculations could be done to report the level of type-II error in the study results.

The SSC for non-comparative studies can be somewhat simpler than the processes typical of sample size computation in comparative studies, since non-comparative studies commonly target an estimate of some quantity of interest, and the SSC is based on achieving a sufficiently precise confidence interval. On the other hand, SSC based on confidence interval width still requires determination of the type of outcome variable, the margin of error (also called precision), the sample variance, and the confidence level (usually 95%). The desired margin of error (half of the desired confidence interval width) should be selected to ensure that the resulting estimate can be interpreted meaningfully. The variance reflects how spread out a variable is, i.e., the variability or dispersion around its central value, this being a mean or a frequency. Estimating the variance of the primary endpoint is critical, as it directly influences the final sample size, because the larger the variance, the larger the sample. The variance of the outcome variable can be estimated by searching similar studies in the literature or pilot study results, or, if these are not available, mathematical models can be applied.

Let's try an example using the formula below. An investigator plans to assess the RR of a chemotherapy regimen in patients treated in the community. She suspects the RR is similar to that reported by clinical trials, albeit it could be slightly lower because cancer patients treated outside of clinical trials tend to have more comorbid illnesses, more commonly are of ECOG 2 status, and are older, etc.; these factors may impact treatment adherence, dose-intensity, and ultimately the treatment outcomes. How many patients (*n*) are necessary to answer this research question? First we need to determine what would be an acceptable RR for the community patients; considering that the RR for the given regimen as reported by phase III trials is 40–45%, it would satisfactory if off-trial patients experienced an RR of at least 35%

(P). Conventionally, an acceptable margin of error (d) for frequencies is ±5%; the narrower the margin, the larger the sample. z statistics indicate the confidence that researchers want to introduce into the SSC. The z statistic is the number of standard deviations by which a standard normal observation is above or below the population mean [21]. For moderately large samples, the sampling distribution of most estimators is approximately normal due to generalizations of the central limit theorem (CLT) (see Chap. 3). The CLT states that as a sample size gets larger, the sampling distribution of the mean for any random variable will approach a normal distribution, irrespective of whether the initial distribution was skewed. The CLT applies to continuous, binary, and time-to-event variables.

For a 95% confidence interval, the z value is 1.96, because 95% of a normal distribution lies within 1.96 standard deviations on each side of the mean. Using the formula below, we come up with $n = 350$, which means that a sample of 350 patients is adequate for assessing an RR of 35% ±5% with a 95% confidence interval; assuming 10% missing data, the final sample can be 385.

$$n = \frac{\left(z^2\right) \times P \times \left(1 - P\right)}{d^2} \quad \frac{350 = \left(1.96^2\right) \times 0.35 \times \left(1 - 0.35\right)}{0.05^2}$$

Prognostic or predictive models also have sample size considerations. For multivariable models, a common rule of thumb is that there should be at least ten observations for each independent variable contained in the model; this is to avoid unreliable and ungeneralizable findings. For example, if researchers want to perform a retrospective cohort study to evaluate prognostic factors associated with worse survival among patients with metastatic choroid melanoma, from how many patients do they need to collect data? If they test nine prognostic factors, each one being a binary or continuous independent variable, then the sample should be at least 90 patients; considering 10–20% of data is missing from medical charts, the final sample is between 100 and 113 patients (100 × 0.9 = 90 and 113 × 0.8 = 90).

Likewise, the dropout rate, and the corresponding missing data rate, in non-comparative studies tends to be higher than these rates in interventional studies, so 10–20% might be expected. It is important to highlight that a well-planned SSC for a non-comparative study does *not* exclude inherent biases associated with uncontrolled studies, such as selection, lead-time, and observational biases.

5.4.2 The Importance of Sample Size Calculation

SSC is at the core of a clinical trial. Without proper SSC, the results of a clinical study can be misleading, not generalizable to other settings, more likely to be false negative or false positive, or unreliable. Hence, not only are methodological problems associated with flawed SSC but there are also ethical implications, such as assumptions of big differences between study groups, which may lead to a lack of statistical power to detect smaller, but still clinically relevant, differences [22]; the planning of oversized trials that enroll an unnecessarily large number of patients to achieve clinically irrelevant results, with consequent waste of time and resources [23]; and undertaking the SSC *after*, rather than before, study initiation. Given all these issues, the detailed reporting of parameters utilized for SSC provides transparency and trustworthiness in study results. Indeed, the Consolidated Standards of Reporting Trials (CONSORT) [24] considers SSC to be a compulsory item in published randomized clinical trials. Yet recent studies have shown that SSC assumptions are poorly reported by clinical trials in oncology. In a cross-sectional survey of 140 phase III trials published in top oncology journals, conducted by our group, we observed that only 27.9% of trials provided all parameters used for SSC [4]. While more than 90% of trials provided the alpha and beta errors, only 57.9% provided data on the expected outcomes of the control and experimental groups, and nearly 20% of phase III trials did not report the planned number of patients to be enrolled.

Conclusions

Clinical research is fundamental for advancing the medical care of patients. And so such research should be based on well-designed studies. Among all the methodological factors to be considered in a study design, sample size is probably the most critical because it directly affects the study results. Meticulous planning and proper reporting of SSC ensures transparency, reliability, and allows reproducibility of results. Additionally, the assumptions for SSC have ethical implications, in that oversized trials may treat an unnecessarily large number of participants, while underpowered trials may be wasteful when they lead to false-negative results. The proper planning for SSC requires time and consideration of the statistical parameters. For the most basic SSC, readily available online free software can be trusted. However, while this chapter was envisioned to help clinical researchers to compute the sample size for their own studies, we strongly recommend having an experienced statistician on board from study conception.

References

1. Clark TG, Bradburn MJ, Love SB, et al. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89:232–8.
2. Guller U, Oertli D. Sample size matters: a guide for surgeons. World J Surg. 2005;29(5):601.
3. Dubey SD. Some thoughts on the one-sided and two-sided tests. J Biopharm Stat. 1991;1:139–50.
4. Bariani GM, de Celis Ferrari AC, Precivale M, et al. Sample size calculation in oncology trials: quality of reporting and implications for clinical cancer research. Am J Clin Oncol. 2015;38:570–4.
5. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982;38:143–51.
6. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis. 1961;13:346–53.
7. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10:1–10.

8. Saad ED, Sasse EC, Borghesi G, et al. Formal statistical testing and inference in randomized phase II trials in medical oncology. Am J Clin Oncol. 2013;36:143–5.
9. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treat Rep. 1985;69:1375–81.
10. Waddell T, Chau I, Cunningham D, et al. Epirubicin, oxaliplatin, and capecitabine with or without panitumumab for patients with previously untreated advanced oesophagogastric cancer (REAL3): a randomised, open-label phase 3 trial. Lancet Oncol. 2013;14:481–9.
11. Riechelmann RP, Alex A, Cruz L, et al. Non-inferiority cancer clinical trials: scope and purposes underlying their design. Ann Oncol. 2013;24(7):1942.
12. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med. 2011;26:192–6.
13. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol. 2003;3:26.
14. Green S, Liu PY, O'Sullivan J. Factorial design considerations. J Clin Oncol. 2002;20:3424–30.
15. Bosset JF, Collette L, Calais G, et al. Chemotherapy with preoperative radiotherapy in rectal cancer. N Engl J Med. 2006;355:1114–23.
16. Cunningham D, Starling N, Rao S, et al. Capecitabine and oxaliplatin for advanced esophagogastric cancer. N Engl J Med. 2008;358:36–46.
17. Kernan WN, Viscoli CM, Makuch RW, et al. Stratified randomization for clinical trials. J Clin Epidemiol. 1999;52:19–26.
18. Mayer RJ, Van Cutsem E, Falcone A, et al. Randomized trial of TAS-102 for refractory metastatic colorectal cancer. N Engl J Med. 2015;372:1909–19.
19. Green SJ, Fleming TR, O'Fallon JR. Policies for study monitoring and interim reporting of results. J Clin Oncol. 1987;5:1477–84.
20. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. Stat Med. 2003;22:953–69.
21. Arya R, Antonisamy B, Kumar S. Sample size estimation in prevalence studies. Indian J Pediatr. 2012;79:1482–8.
22. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA. 2002;288:358–62.
23. Altman DG. Statistics and ethics in medical research: III How large a sample? Br Med J. 1980;281:1336–8.
24. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. Open Med. 2010;4:e60–8.

Interpretation of Results from Tables, Graphs, and Regressions in Cancer Research

6

Raphael L. C. Araújo and Rachel P. Riechelmann

6.1 Introduction

Tables and graphs are often used in scientific articles to summarize the study results for readers. Although the concept seems to be pretty simple, and it is, it may be difficult for researchers to create Tables and graphs in appropriate ways, and it may be difficult for readers who are not familiar with clinical research to interpret the information in the Tables and graphs. Here we present the characteristics of Tables, graphs, Figures, and schemes that are most commonly utilized in cancer clinical research; as well, we present discussions and recommendations about their interpretation.

Tables and graphs are very useful forms for the reporting of univariate and multivariate analyses, which are the usual way of identifying associations between exposure and outcomes. This chapter highlights the rational use and interpretation of Tables, graphs, curves, and summarises results of regression analyses. Linear, logistic, and Cox regressions are discussed in regard to their interpretations, main advantages, and limitations.

6.2 Tables

The main objectives of creating a Table are to summarize the information gained from the study in a fashion that is clear and accessible for readers. Most studies produce many results with numerous variables; if all results are described in the

R. L. C. Araújo, M.D., Ph.D. (✉)
Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos Cancer Hospital, Barretos, SP, Brazil

R. P. Riechelmann, M.D., Ph.D. (✉)
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_6

text, the manuscript will be too detailed and too long, and relevant scientific information could be diluted among so many findings. On the other hand, not all data needs to be shown in Tables. While there are no fixed rules about when to create a Table for reporting study results, it is common sense that patient characteristics and outcomes with very many variables and results should be presented in Tables.

Most clinical research articles have a Table in which a summary of the population characteristics is presented; this is often presented in absolute numbers, percentages, and/or medians/means, with the relevant measures of dispersion. For example, in the following sentence it is clear that a Table (see Table 6.1) was used to compile the detailed information about the study population: "*Clinicopathological and operative data are outlined in Table 6.1. The population studied included 401 (49.0%) female and 418 (51.0%) male patients with a mean age at the date of operation of 66 ± 12 years (range: 29–92 years; median: 67 years). Significant*

Table 6.1 Clinicopathological and operative variables in patients who underwent pancreaticoduodenectomies. Extracted from Araujo RL et al. [1]

		Complications		
Variable	All patients (n = 819)	Yes (n = 405, 49.5%)	No (n = 414, 50.5%)	P-value
Age, years, median (IQR)	67 (58–75)	67 (59–76)	68 (57–75)	0.638
Sex, *n* (%)				<0.001
Female	401 (49.0%)	171 (42.2%)	230 (55.5%)	
Male	418 (51.0%)	234 (57.8%)	184 (44.5%)	
Body mass index, kg/m^2, median (IQR)	26.4 (23.4–29.7)	27.0 (23.8–30.5)	25.8 (22.8–29.2)	0.002
Hypertension, *n* (%)	392 (47.9%)	199 (49.1%)	193 (46.6%)	0.485
Diabetes mellitus, *n* (%)	157 (19.2%)	76 (18.8%)	81 (19.6%)	0.790
Cardiac disease, *n* (%)	189 (23.1%)	107 (26.4%)	82 (19.8%)	0.025
Pulmonary disease, *n* (%)	80 (9.8%)	47 (11.6%)	33 (8.0%)	0.099
Other comorbidities, *n* (%)	533 (65.1%)	258 (63.7%)	275 (66.4%)	0.421
ASA physical status, *n* (%)				0.106
Class 1	17 (2.1%)	6 (1.5%)	11 (2.7%)	
Class 2	422 (51.5%)	201 (49.6%)	221 (53.4%)	
Class 3	370 (45.2%)	190 (46.9%)	181 (43.7%)	
Class 4	10 (1.2%)	8 (2.0%)	2 (0.5%)	
Greatest diameter of tumour, cm, median (IQR)	782 (95.5%)[a]	2.8 (2.0–3.6)	3.0 (2.0–3.8)	0.140
Malignant tumours[b], *n* (%)	701 (85.6%)	342 (48.8%)	359 (51.2%)	0.319
Diagnosis, *n* (%)				
Ampulla of Vater				
Adenocarcinoma	139 (17.0%)	74 (18.3%)	65 (15.7%)	
Adenoma	8 (1.0%)	3 (0.7%)	5 (1.2%)	
Neuroendocrine tumour	3 (0.4%)	1 (0.2%)	2 (0.5%)	
Others	4 (0.5%)	2 (0.5%)	2 (0.5%)	
Bile duct				
Adenocarcinoma	55 (6.7%)	33 (8.1%)	22 (5.3%)	

(continued)

		Complications		
Variable	All patients (n = 819)	Yes (n = 405, 49.5%)	No (n = 414, 50.5%)	P-value
Neuroendocrine tumour	1 (0.1%)	1 (0.2%)	0	
Others	7 (0.9%)	5 (1.2%)	2 (0.5%)	
Duodenum				
Adenocarcinoma	55 (6.7%)	28 (6.9%)	27 (6.5%)	
Adenoma	9 (1.1%)	5 (1.2%)	4 (1.0%)	
GIST	6 (0.7%)	5 (1.2%)	1 (0.2%)	
Neuroendocrine tumour	4 (0.5%)	2 (0.5%)	2 (0.5%)	
Others	1 (0.1%)	1 (0.2%)	0	
Pancreas				
Adenocarcinoma	437 (53.4%)	197 (48.6%)	240 (58.0%)	
Cystadenoma	23 (2.8%)	14 (3.5%)	9 (2.2%)	
Neuroendocrine tumour	36 (4.4%)	22 (5.4%)	14 (3.4%)	
Pancreatitis	19 (2.3%)	7 (1.7%)	12 (2.9%)	
Others	12 (1.5%)	5 (1.2%)	7 (1.7%)	
Procedure, n (%)				
Standard PD	689 (84.1%)	343 (49.8%)	346 (50.2%)	0.702
Pylorus-preserving PD	130 (15.9%)	62 (47.7%)	68 (52.3%)	
Duration of surgery, min, median (IQR)	266 (221–322)	276 (226–330)	261 (217–313)	0.005
Estimated blood loss, ml, median (IQR)	600 (400–900)	600 (400–1000)	500 (350–800)	0.001
Length of stay, days, median (IQR)	9 (8–13)	12 (9–18)	9 (7–10)	
Any positive margins, n (%)	135 (16.5%)	58 (14.3%)	77 (18.6%)	0.109

ASA American Society of Anesthesiologists, *GIST* gastrointestinal stromal tumour, *IQR* interquartile range, *PD* pancreaticoduodenectomy

[a]The total number of the measureable tumours was 782

[b]Malignant tumours were confirmed by pathology in 701 patients

differences in the presence of comorbidities between the groups with and without complications emerged only for cardiac disease (26.4% versus 19.8%, respectively; p = 0.025) and BMI (median: 27 kg/m^2 versus 26 kg/ m^2, respectively; p = 0.002); no significant differences were noted for other comorbidities" [1].

Other Tables can be used to summarize the results of secondary endpoints; for example, frequencies of adverse events, exploratory subgroup analyses, and descriptive data, such as resulting scores for overall health-related quality of life and its domains. The most important result, i.e., the primary endpoint, should be stated in the text but not necessarily in Tables or Figures. For example, survival analyses are generally reported in the text and as Kaplan-Meier estimate curves. If the study's primary endpoint is the objective response rate, for instance, the results could be reported only in the text; but in this case, if there are multiple comparisons of response rates across many subgroups, then a Table may be helpful in presenting the results.

Tables should be self-explanatory, meaning that all the information presented should be clear enough to allow readers to understand the content without much effort.

So the first, and one of most important steps in creating a Table is the choice of the title. The title must be clear enough to allow a reader to interpret it without having to previously read the text. For example, in Table 6.1, all information about the population studied is given in the title, the variables, and the footnotes. The information regarding the question addressed, i.e., the presence or not of postoperative complications in patients who underwent pancreaticoduodenectomies, is contained in the subheading or in the column heads. All variables are seen separately in rows.

Another point that makes information in a Table self-explanatory is to state abbreviations and their meanings in the Table footnotes. Of note, missing data are common in clinical research, especially in retrospective studies. Thus, the population used to measure associations between the groups evaluated can vary according to the variable analyzed. In the Table 6.1 example, differences in data distributions are described in footnotes. The observed differences allow readers to realize how much missing data is present for each variable, and to determine whether the denominators of interest are fair enough to indicate a valid association or not. Regarding multivariate models, the description of missing data is even more important, because statistical software considers a minimal number of variables that are not missing in order to run the analysis with all variables.

Tables are not used exclusively for results; sometimes a Table can present steps in the study methodology, such as selection criteria; steps in experimental procedures; or any other information that would fit better in a Table than in the text. Systematic reviews are a special situation where many different datasets from different articles have to be clearly organized in Tables. In these reviews, the study "subjects" are often articles, not individuals, and Tables assume an essential role in structuring the information, as shown in Table 6.2, which was designed to compile and present the methodology for the quality assessment of the eligible randomized trials. This information certainly could not be reported in the text, because it would be confusing, too detailed, and tedious for readers.

The layout of Tables is important in making all the information clear. Font type and size, lines, outlines, and color cells are keys to achieving a good design. The Table should be as concise as possible, so only pertinent information should be addressed.

Table 6.2 Quality assessment of selected randomized clinical trials in patients who underwent curative-intent treatment for colorectal liver metastases. Extracted from Araujo RL et al. [2]

	Studies		
	Langer [3]	Portier [4]	Nordlinger [5, 6]
Randomized clinical trials evaluated by Cochrane Risk-of-Bias	Tool		
Random sequence generation	Unclear	Low	Low
Allocation concealments	Low	Low	Low
Blinding of participants and personnel[a]	Low	Low	Unclear
Blinding of outcome assessment[a,b]	Low	Low	Low
Incomplete outcome data	Unclear	Low	Low
Selective reporting	Low	Low	Low
Other bias	Unclear	Low	Low

	Studies			
	Parks	Adam [7]	Reddy [8]	Ihemelandu [9]
MINORS score for eligible observational comparative studies				
Clearly stated aim	2	2	2	2
Consecutive patients	2	2	2	0
Prospective data collection	2	2	1	2
Appropriate end points	2	2	1	2
Unbiased outcome evaluation	1	1	0	0
Appropriate follow-up	1	1	0	1
Loss to follow-up ≤5%	0	1	0	0
Prospective calculation of study size	0	0	0	0
Adequate control group	2	2	2	2
Contemporary groups	2	2	2	1
Baseline equivalence of groups	1	1	1	1
Adequate statistical analysis	2	2	1	2
MINORS score index	17	18	12	13

MINORS Methodological Index for Nonrandomized Studies
[a]Blinding is not possible
[b]Implementation of a protocol for postoperative management was considered the best alternative

6.3 Graphs

Graphs are pictures that represent how variables are related. Many kinds of graphs are used in clinical research, and they vary according to the type of data and analyses used: data distribution, variation over time, and comparison between variables. Like Tables, graphs should also be self-explanatory, without a reader needing to read the whole text to understand what the graphs mean. So titles and legends are crucial to facilitate the complete understanding of the depicted information. Regarding data distribution, the main graphs used are histograms, pie charts, bar charts, box plots, and forest plots. Although these graphs are often used for one variable, they can be useful for subgroup analysis or combined analysis as well.

6.3.1 Histograms

Histograms are the graphs usually chosen to represent classes of a continuous variable or some discrete variables (ordinal variables). The range of the studied variable is placed on the horizontal axis and the frequency, on the vertical axis, as demonstrated in Fig. 6.1a, b. Both parts a and b of this figure represent distributions of continuous variables (age and weight) in a hypothetical population of 120 patients who underwent treatment for liver cancer [10]. In both graphs, the y-axis (vertical) represents frequency, and each bar presents the total number of observations that correspond to the column represented on the x-axis (horizontal). There is no need to always present both total number and frequency in the graph; they are shown in the figure to emphasize that the y-axis represents frequency, and not absolute numbers. Another important consideration is that both graphs present lines following the data

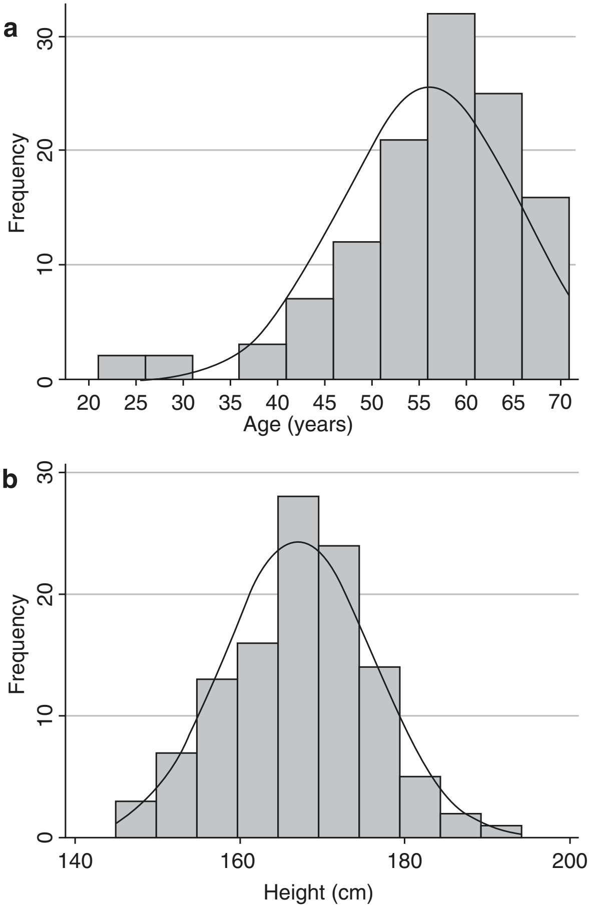


Fig. 6.1 Histograms representing distributions, according to age (**a**) and height (**b**), of a hypothetical population of 120 patients who underwent liver cancer treatment

distribution. These lines are different; the age density line (Fig. 6.1a) does not fit as normal as the height line (Fig. 6.1b). In doubtful cases, as for these lines, some tests of normality can be applied to verify whether there is a non-normal distribution. Here, the Shapiro-Wilk test was applied; it did not demonstrate normality in Fig. 6.1 (age), with $p < 0.001$, but it demonstrated normality for the histogram in Fig. 6.1b (height), with $p = 0.766$. Normal and non-normal distributions are described in more detail in Chap. 3.

6.3.2 Pie Charts

Distributions of categorical variables are often presented in pie charts. Figure 6.2 shows distributions according to blood types in the same hypothetical population of patients as that shown in Fig. 6.1. The respective proportions are presented as slices, and the whole pie corresponds to the whole sample.

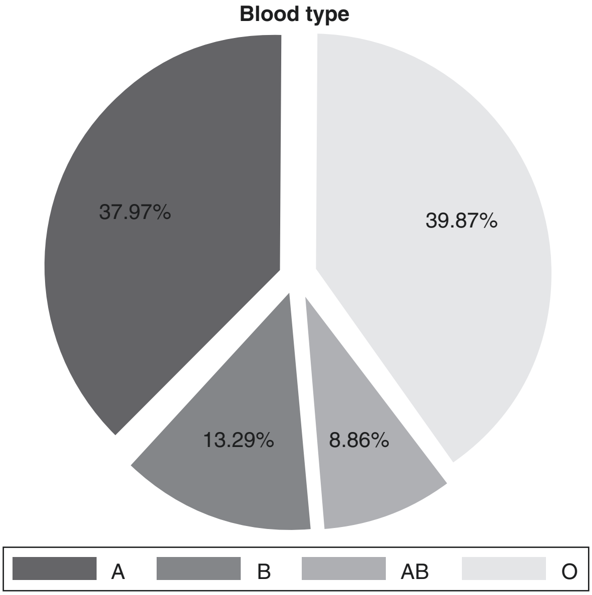


Fig. 6.2 Pie chart representing distributions, according to blood type, of the hypothetical population of 120 patients who underwent liver cancer treatment

6.3.3 Bar Graphs

Bar graphs resemble parallel bars, with the same width, and commonly compare quantities or proportions of different categories or groups. All bars are put in parallel on one axis and the unit of measure is stated on the other axis. Both lines and bars can have the same purpose, e.g., comparing progress in different categories and frequency in different groups, as depicted in Fig. 6.3a, b. Lines and bars can represent two different ways of illustrating the same idea, as shown in the graph of complications according to the month of the year (Fig. 6.3a): the total number of surgeries and different categories of complications are shown. It is also possible to use bars grouped in the same column to show proportions; thus, the whole column represents 100%, and each part of the column represents the respective subgroup (Fig. 6.3b).

6.3.4 Box Plots

Box plots are useful to show data distribution for continuous variables, with or without comparison between groups. The concept of the box plot is to represent medians, percentiles, and outliers of continuous data, as depicted in Fig. 6.4a, b. Much information is contained in the box plot graphs. The box properly represents interquartile ranges (25th to 75th percentiles), i.e., the most representative part of the study population; in fact, half of it (50%). The other half of the population is split into the under 25th percentile (lower hinge) or the over 75th percentile (upper hinge). Although these are both parts of the whole population, they are not representative of the bulk of the sample because they contain outliers. Generally, box plots present continuous variables on the vertical axis, and groups or categories on the

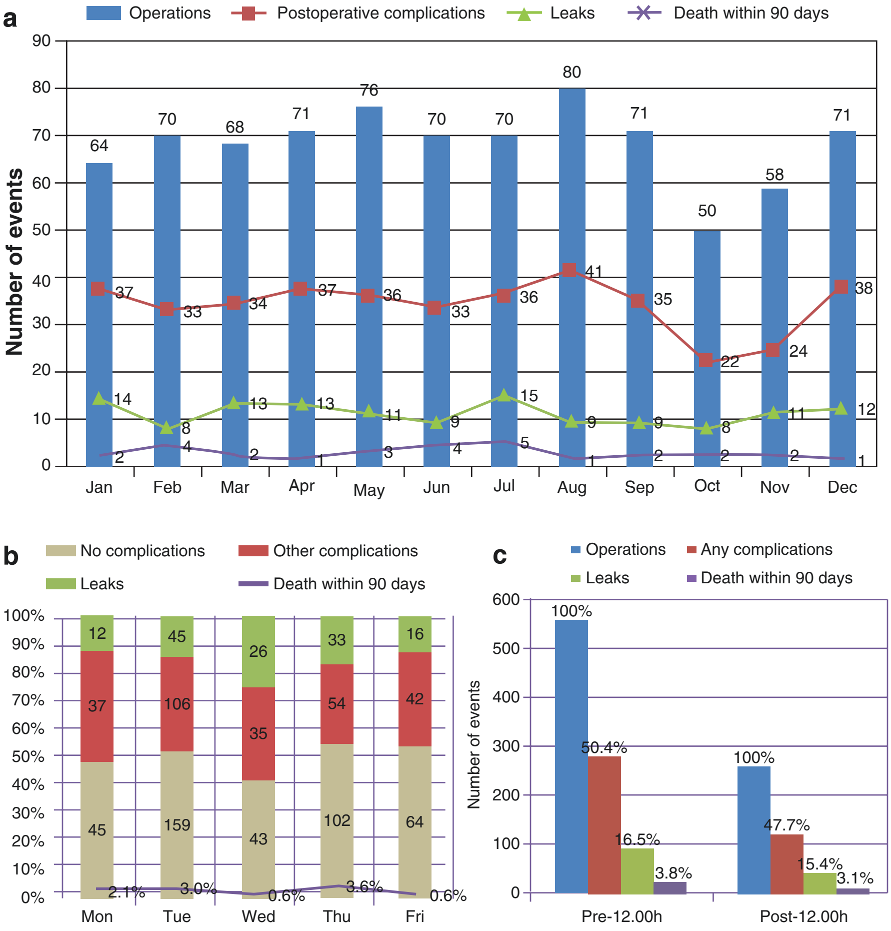


Fig. 6.3 Distribution of 819 pancreaticoduodenectomies according to surgical scheduling in the 9-year period studied (2000–2008). (**a**) Distribution of operations according to related postoperative complications, leaks, and 90-day mortality stratified by month (P = 0.920, P = 0.715, and P = 0.736, respectively). (**b**) Frequency of postoperative complications, leaks and fistulae, and 90-day mortality stratified by day of the week (P = 0.279, P = 0.097, and P = 0.114, respectively). (**c**) Number of operations associated with complications, leaks and fistulae, and 90-day mortality stratified by start time (prior to or after 12.00 h) (P = 0.057, P = 0.760, and P = 0.690, respectively). From Araujo RL et al. [1]

horizontal axis. But it is possible to have box plots with horizontal position just pivoting numerical and categorical variable as well. Another remarkable point in box plot analysis is that larger boxes indicate smaller samples. In Fig. 6.4b, the box for the male group is bigger than the box for the female group, and the reason is that the samples contain 90 (75%) patients versus 30 (25%), respectively. A larger sample is associated with smaller variance, and consequently, a smaller confidence interval (CI) and narrower interquartile range; also, a larger sample is represented by a smaller box than the one that represents a smaller sample.

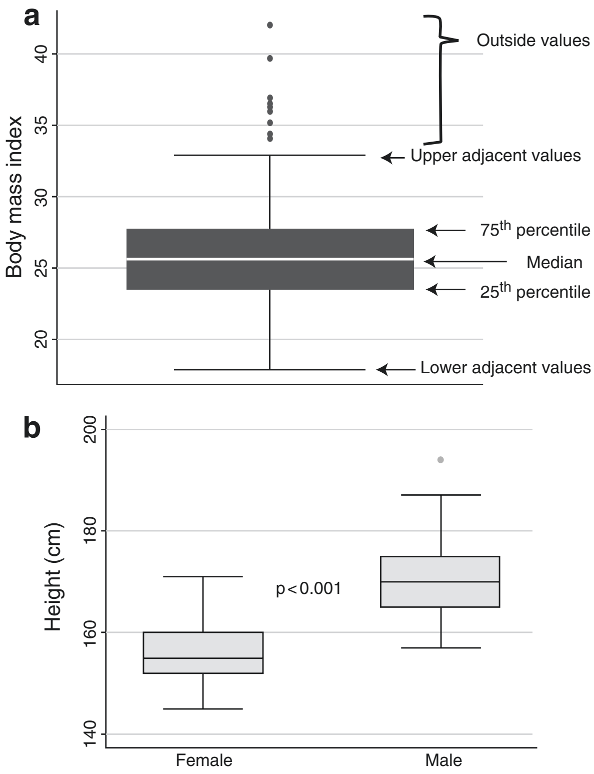


Fig. 6.4 Hypothetical and practical examples of box plots. (**a**) Represents the elements of box plots. (**b**) Represents the distribution, according to sex, of a hypothetical sample of 120 patients who underwent liver cancer treatment

6.3.5 Forest Plots

Another graph that is useful in oncologic studies is the forest plot. Its importance increases in systematic reviews with meta-analysis and also with the analysis of subgroups in clinical trials, for example. The role of forest plots in meta-analysis is to summarize and report the overall results of the pooled data extracted from the published papers included in the systematic review. Indeed, forest plots demonstrate the effect size (or magnitude of difference) of the comparison between two different therapeutic interventions. In a forest plot (Fig. 6.5), the horizontal axis presents the effect sizes of the outcomes reported by studies, which is the primary endpoint of the meta-analysis. The outcomes are generally based on ratios such as the odds ratio (OR), relative risk (RR), or hazard ratio (HR). In Fig. 6.5, the vertical axis intersects with mark one (1) on the horizontal axis, and this position represents the null effect. It splits the graph area into two opposite sides according to the effect size of each study in regard to the control arm versus the intervention arm. The effect size of each article included, as well as the pooled analysis, demonstrates whether the results favor the control or the experimental treatment.

Another important point in the forest plot is the square, as it represents the effect estimate; its size varies according to the weight (or size) of the sample. When

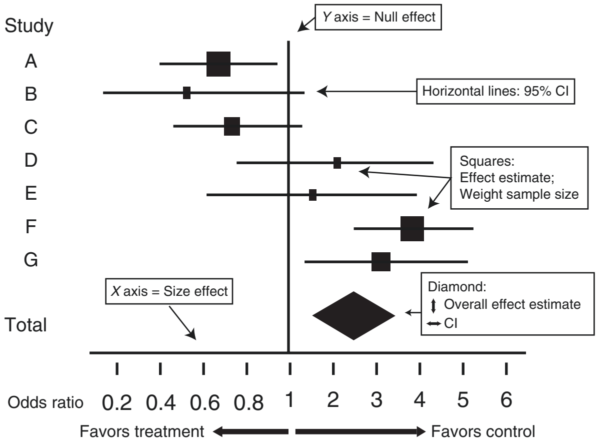


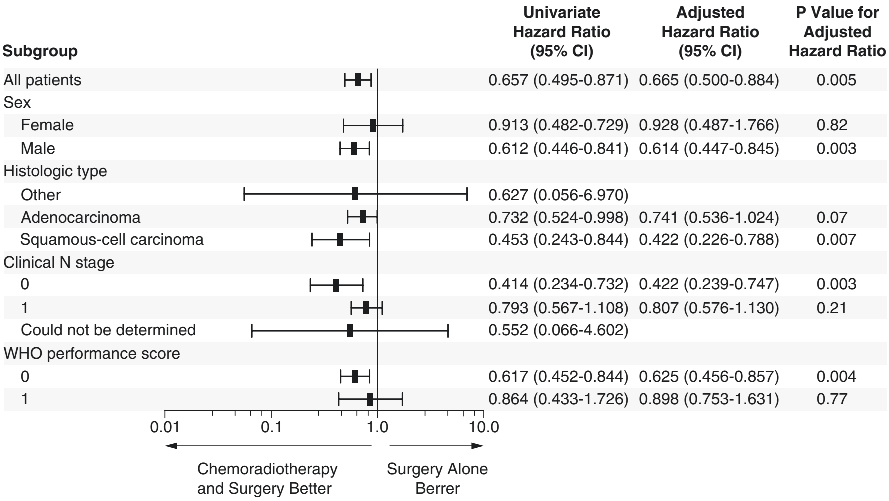
Fig. 6.5 Hypothetical forest plot. *CI* confidence interval

squares show ratios or risk over 1, i.e., to the left of the vertical axis, they demonstrate an increasing likelihood that the experimental treatment group has a more favorable outcome. If the squares are plotted to the right of the vertical axis, this means there is a higher probability of events in the control arm. However, squares must not be analyzed alone, because there is another element crucial to forest plot interpretation; i.e., the horizontal lines. These lines represent the 95% CI. If the CI line touches the y-axis, this means there is no significant difference between the groups. If the line is entirely on one of the sides of the null effect line (y line), this means that the estimated effect is favorable for this respective arm. Finally, the diamond in a forest plot represents the averaged result of all size effects and CI combined. As expected, the larger the sample, the narrower and more precise is the CI and the horizontal size of the diamond (systematic reviews and meta-analyses are discussed in detail in Chap. 18).

An alternative use of forest plots is for the analysis of subgroups. For instance, during a randomized clinical trial, a subset of patients can present different treatment outcomes, as demonstrated by van Hagen et al. in a phase III randomized clinical trial of preoperative chemoradiation with carboplatin and paclitaxel and surgery versus surgery alone for esophageal or junctional cancer [11]. Patients were analyzed according to their histologic types; differences in the HR for deaths are clearly depicted in the forest plot, as demonstrated in Fig. 6.6.

6.4 Regressions

Tables and graphs for regressions are described separately in this chapter because regressions have a notable importance in cancer research. Various regression models are widely employed in clinical research, since they provide statistical approaches for estimating the risks of events associated with various characteristics presented by study participants. Regression analysis can be done as simple regression (univariate regression), relating a unique variable to the event of interest. This analysis



Subgroup	Univariate Hazard Ratio (95% CI)	Adjusted Hazard Ratio (95% CI)	P Value for Adjusted Hazard Ratio
All patients	0.657 (0.495-0.871)	0.665 (0.500-0.884)	0.005
Sex			
Female	0.913 (0.482-0.729)	0.928 (0.487-1.766)	0.82
Male	0.612 (0.446-0.841)	0.614 (0.447-0.845)	0.003
Histologic type			
Other	0.627 (0.056-6.970)		
Adenocarcinoma	0.732 (0.524-0.998)	0.741 (0.536-1.024)	0.07
Squamous-cell carcinoma	0.453 (0.243-0.844)	0.422 (0.226-0.788)	0.007
Clinical N stage			
0	0.414 (0.234-0.732)	0.422 (0.239-0.747)	0.003
1	0.793 (0.567-1.108)	0.807 (0.576-1.130)	0.21
Could not be determined	0.552 (0.066-4.602)		
WHO performance score			
0	0.617 (0.452-0.844)	0.625 (0.456-0.857)	0.004
1	0.864 (0.433-1.726)	0.898 (0.753-1.631)	0.77

Fig. 6.6 This forest plot shows hazard ratios for death, and 95% confidence intervals, for 366 patients with esophageal or esophagogastric-junction cancer, according to their baseline characteristics. From van Hagen et al. [11]

is often used in clinical research, since it results in a mathematical function describing outcomes according to the presence or not of a predictive variable. If more than one variable is assessed in the model (multiple variables), a multivariate analysis is performed, with the aim being to find how and to what extent each variable would contribute to the final outcome under evaluation.

As we have different types of variables, it is expected that we would also have different types of regression analyses. If the variable assessed is of the continuous type, it needs a linear regression model to describe a linear function, although more complex non-linear models may be used. Binominal variables are evaluated using a logistic regression, and time-to-event variables are assessed using Cox regression analysis. As well as being used in clinical trials, regression analyses can be used in observational studies, where they aim to control for confounding factors that could influence the primary endpoint result. The three types of regression analyses will be discussed separately below, and their clinical use in clinical cancer research will also be covered.

In retrospective series, regressions are important to identify the strength of the association between exposure and outcomes. Regressions are useful to show the benefits and harms of some practices, such as, for example, showing predictors of severe toxicity. Multivariate analyses are also useful to identify confounding variables that are falsely associated with the investigated outcomes. In randomized controlled clinical trials, variables tend to be balanced between groups because of the randomization; however, some imbalance can still happen simply due to chance. In such cases, regression analyses adjust the overall result of the primary endpoint to covariates that might have influenced the result. These analyses suggest a cause-effect relationship, which is not always easy to determine, since many uncontrolled (and unknown) variables may interfere with results. More importantly, both univariate and multivariate regression analyses rely on variables that were selected by investigators. Thus, the selection process of picking the most scientifically relevant variable is crucial for study design and inferences of causality. Finally, while regression analyses can answer many questions in clinical trials, their findings often generate new hypotheses to be tested in future studies.

6.4.1 Linear Regression

Although the name linear regression sounds as if it is a complex statistical tool, it could be more easily interpreted if it were regarded as a linear function: $y = ax + b$, the same one that is usually presented in basic mathematics. This simple linear regression predicts a value according to another value. What is still unknown, and which needs to be predicted by the formula, is the criterion variable *y,* also called a dependent variable. The variable that is already known is named *x,* and is already defined; therefore it is called an independent variable. Hence, linear regression is no more than a simple function, with clinical variables.

Here we present a clinical oncology example of a hypothetical population of 120 patients who underwent liver cancer treatment. It is reasonable to think that the number of lesions and their size could be associated; to demonstrate their association, a simple linear regression can be useful. In this example, the median number of lesions was 1 (1–3 interquartile range) and the median size of the largest lesion was 24 mm (17.5–30 interquartile range). The number of lesions (*y*) is what we want to estimate, the size of lesions is *x,* and *a* and *b* are called coefficients or constants:

$$y = ax + b$$

Looking back to the algebraic world, *y*, the dependent variable, is what we want to predict and it will come up from the equation; *x*, the independent variable, is what we independently know; *a* is the constant governing the slope of the linear function demonstrated in the fitted line; and *b* represents the constant where the line crosses the y axis, the y-intercept point. Regarding the equation below, if *x* was zero, *y* would be 1.32, as suggested by the dashed line in Fig. 6.7.

$$y = (0.045) \times (x) + 1.32$$
$$\text{Number of lesions} = (0.045) \times (\text{size of lesions}) + 1.32;\ \text{CI 95\% } (0.02 - 0.07);\ p = 0.001$$

The dashed line in Fig. 6.7 represents a linear fit adjusting a line that graphically shows the linear prediction built on this linear regression. The idea of the line is to fit the values in a single line; thus adapting them to a linear prediction made by regression. It is remarkable that the depicted outliers can be adjusted to a linear fit, since it represents the predictions generated by the formula. The line shows the mathematical fitting of all values, considering that all of them are adjusted by the same behavior.

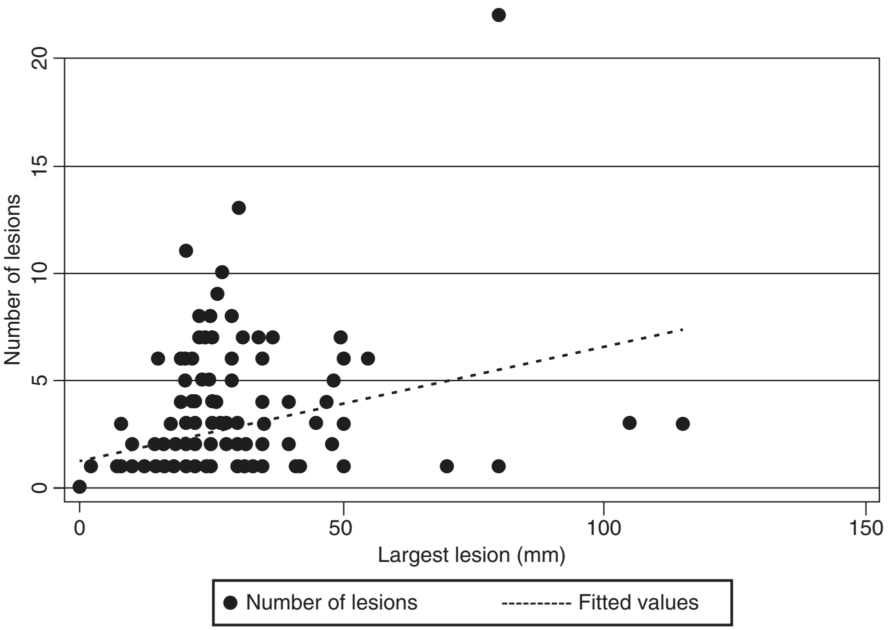


Fig. 6.7 Linear regression of the number of lesions, according to the size of the largest lesions, for a hypothetical population of 120 patients with liver cancer

6.4.2 Logistic Regression

Logistic regression is a very useful tool in clinical cancer research because many clinical questions are dichotomous (also called binary). Logistic regression analyses can address clinical inquiries, such as, for example, the presence of complications, diseases, adverse events, etc. Logistic regressions can also look for a dichotomous variable versus a continuous variable. Simple logistic regression implies the prediction of a dichotomous result (yes or no) according to a continuous independent variable, or the probability of an event of interest (yes or no) occurring according to the score of a continuous variable.

Logistic regression should not be used to determine a diagnosis such as cancer, or to determine complications or any other outcome. It is important to understand that logistic regression attempts to predict the probability of an event of interest based only on the independent variable. The correct choice of the independent clinical variable is therefore crucial to achieve clinically relevant results. This choice is the responsibility of the investigators; statisticians can help with numeric analyses and find associations, but it is not their role to decide which variables should be tested in a model.

The results of logistic regressions are expressed as odds ratios (ORs), i.e., a measure of association between independent (exposure) and dependent (outcomes) variables. The OR represents the chance that the events would occur given a specific exposure, that there is no difference in outcomes, OR > 1 indicates that the risk of events is higher for those who are exposed, and OR < 1 means that the risk of events is lower among the exposed individuals [12]. But ORs are not independent scores, they have to be tested for statistical significance, either with *p* values or confidence intervals. Therefore, ORs can help to interpret whether the association observed between exposure and no exposure in the analyzed population was beneficial or detrimental.

6.4.3 Cox Regression (Hazard-Proportional Model)

The analysis of survival, which is of special interest in clinical research, is widely used in oncology research. Although survival seems to be just a dichotomous variable (dead or alive, recurrence or not, etc.), survival analysis is a model that requires the observation of time. Patients should be followed for enough time to allow for the occurrence of the event. The problem is that not all patients experience the event during the follow-up period. An example of this possible drawback is a study that evaluates the oncologic outcomes of patients with indolent tumors whose median study follow-up time is 1 year. This follow-up time would clearly bias the survival rates, showing lower rates of death or recurrence. Diseases with longer median survival, such as metastatic colorectal cancer, require longer follow-up times, since many effective therapeutic options are available, including curative-intent surgical resection for metastatic disease. On the other hand, in diseases with high lethality and shorter overall survival, such as advanced pancreatic cancer, the median follow-up can be shorter, since the events of interest will not take so long to occur. Another important concept in survival analysis is the censoring of data, which represents a patient who is lost to follow-up and in whom the event of interest had not occurred at last contact.

In the hazard proportional model, two variables are necessary to run the model: one is the time of observation and the other is whether the event occurred or the time is censored. All individuals are counted to the final analysis according to their respective times of observation, from the date of the completed treatment (or randomization, treatment start, etc.) until the date the event of interest occurs or the date when the follow-up is censored, whichever comes first. Regarding each observation, the duration of observation is assessed from the starting point (according to the outcome evaluated) until the date the event of interest is recorded (with images, biopsy samples, laboratory findings, or even clinical examination), or the censored date. It is also important to remember that patients are not all enrolled at the same time, so the times of observation vary for each patient, as depicted in Fig. 6.8.

The follow-up time needs to be adjusted according to the duration of a study, as this time impacts the frequency of events of interest. Time variation is demonstrated in Fig. 6.9, with three different observation times exemplified and with different rates for 3-year follow-up (A - 2/8, and B - 4/8). This scheme attempts to show the importance of longer follow-up, especially for overall survival. Thus, a longer follow-up time is essential for achieving good quality data and better survival rates. Figure 6.10 shows a global analysis, using the same

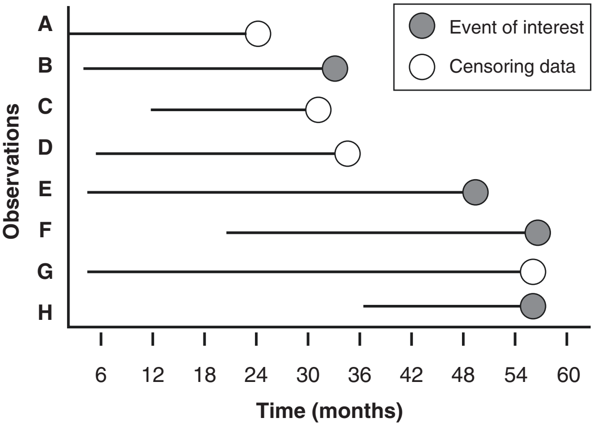


Fig. 6.8 Hypothetical distribution of observations according to patients' times of entry into the study and their follow-up times (for the same population as that in Fig. 6.1)

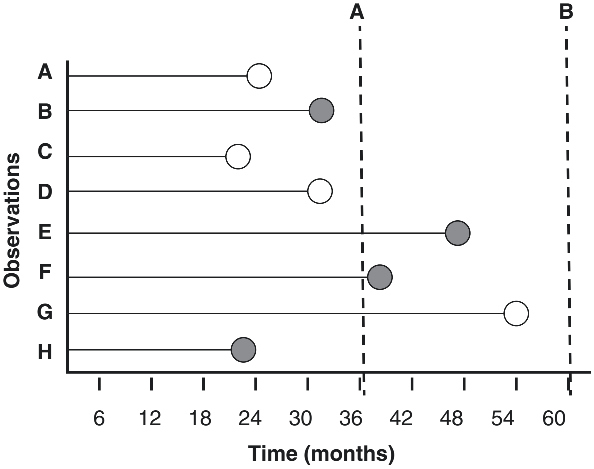


Fig. 6.9 Differences in observations and outcomes related to times of observation (36 and 60 months), for the same population as that in Fig. 6.1

hypothetical database of 120 patients with liver cancer as that shown in Fig. 6.1. It is important to highlight that, as this was a global analysis, no hypotheses or tests were applied.

Comparisons between survival curves are usually performed to compare oncologic treatments or prognostic factors in oncology. Thus, a hypothesis has to be tested, and then the log-rank test is the typical choice. Here we present another example comparing two treatment modalities (perioperative and postoperative chemotherapy) after potentially curative hepatic resection for metastatic colorectal cancer, and their impact on recurrence-free survival, as depicted in Fig. 6.11 [13]. Two curves are shown and the curves are not overlapping. This corroborates the statistical significance demonstrated by the log-rank test ($p = 0.036$).

Regarding analysis with graphs, the Kaplan-Meier estimate is the most frequently used graph, since it gives a clear estimation and compares risks of events over time. More information about survival analysis and censored data is available in Chap. 7.

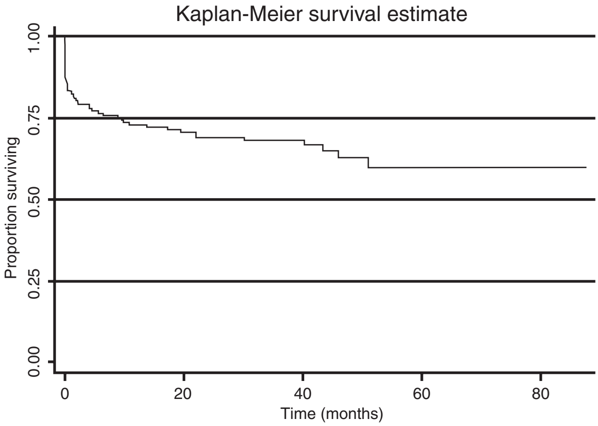


Fig. 6.10 Overall survival of a hypothetical population of 120 patients who underwent liver cancer treatment

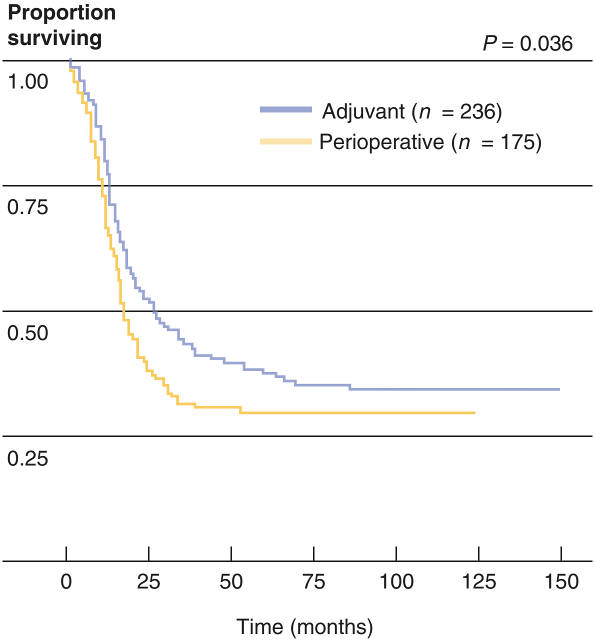


Fig. 6.11 Kaplan-Meier estimates of recurrence-free survival from the date of liver resection, in months, for 411 patients who underwent liver resection for resectable colorectal liver metastases, according to the timing of chemotherapy (adjuvant versus perioperative)

Using the same example as the one above, from Fig. 6.11, it is possible to extract some information from Kaplan-Meier models, as demonstrated in Fig. 6.12, in which the vertical axis represents the probability of survival; this axis starts at 100% of the population and every time that the event of interest occurs, the curves fall in a stepwise fashion, representing a reduction in the population free of recurrence. The horizontal axis represents the time that each individual was followed until the present recurrence, or the last follow-up for those who did not present recurrence. As shown in Fig. 6.7, with longer follow-up, differences became clear with the separation of the two curves. The difference between the curves was investigated by the log-rank non-parametric statistical test, which showed a significant difference ($p = 0.036$).

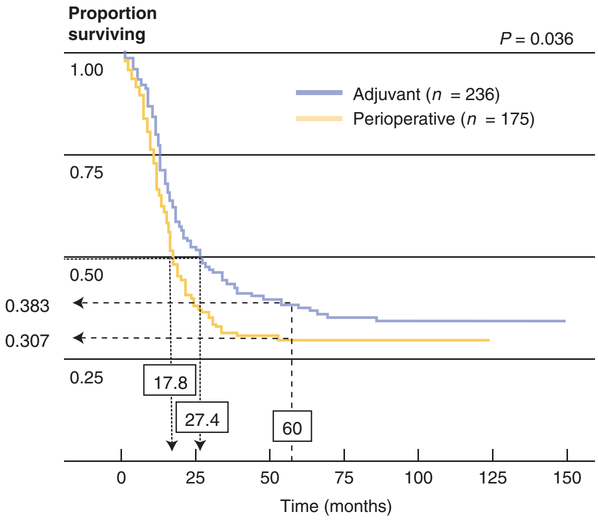


Fig. 6.12 Interpretation of survival data in a Kaplan-Meier graph. The line with large dashes represents different ways of obtaining survival information from the curves. The arrows with small points represent median survival data. Data in the boxes represent months and data on the y-axis represent proportions. The line with small dashes shows 5-year survival

In survival analyses, survival can also be described as median survival and specific survival by time; for instance, the 5-year survival rate. In Fig. 6.8, the point 0.5 on the y-axis demonstrates that half of the population (median) recurred at this plateau, with recurrences for the adjuvant and perioperative groups, respectively, occurring at 17.8 and 27.4 months. So, according to the information on the y-axis, it is possible to link the median recurrence-free survival on the x-axis by using the curves. It is also possible, when choosing a specific time mark (i.e., 5-year survival; 60 months), to identify the proportion of the surviving population on the y-axis.

6.4.4 Univariate Analysis

Univariate analysis is a method of testing the association of one or many individual variables against the independent or outcome variable. The analysis of each variable separately can suggest associations of one or some variables with the event of interest, as, for example, the presence of postoperative complications (yes or no—dichotomous variable). Thus, looking for association between variables, a null hypothesis is tested and statistical tests and a *p* value will be found. The *p* value represents the probability of the null hypotheses being rejected, or merely the probability of that association being not by chance. It is common to consider the *p* value significant when it is ≤0.05.

Two important points about the *p* value merit discussion. There is a logical axiom in statistics that deserves to be quoted: “absence of evidence does not mean evidence of absence”. If the null hypothesis is rejected, the study is significant; however, if the null hypothesis is not rejected, that does not necessarily mean that there is no difference; it simply means that the difference is not statistically significant. The second point is that *p* values cannot be compared as markers of evidence. A *p* value of 0.04 is not less significant than a *p* value of 0.02. Each *p* value addresses a specific question, and these values represent tests of null hypotheses, or tests of association, and not the effect size of a comparison. Thus, it does not make sense to compare *p* values.

Univariate analyses vary according to the event of interest and also according to the independent variables being examined (categorical or dichotomous; continuous; and time-to-event). The main statistical tests used in univariate and multivariate analyses are dealt with separately in Chap. 4 and Chap. 7.

6.4.5 Multivariate Analysis

In univariate analysis, the variables are tested separately to address their individual roles as predictors of the outcome. However, many of these variables can have either a synergic effect between them or they can act as confounding variables in regard to the final outcome. A multivariate model attempts to measure the degree of each variable's contribution to the event of interest in a given population, and to measure the effect of each variable independently when they are evaluated together. Although the idea seems simple, and it is, many tricks and possible drawbacks can be found in multivariate models.

In regard to the variables used for univariate and multivariate analyses, the first concept is to realize that statistical software cannot discern which variables should be included in these analyses, so the most important point for the investigator is the selection of the relevant clinical variables. This selection has to be carefully considered, based on the investigator's clinical judgment. It is also important to emphasize that statistical relevance does not necessarily mean clinical relevance. The second concept is linked to the first, since the selection of variables that should be included in the multivariate analysis originates from the variables tested in the univariate analysis. In the univariate analysis, variables that presented a statistical association with the event of interest are identified, and those variables should ideally be included in the multivariate model. Most studies consider an overall significance level as being lower than 0.05; thus, variables that present such association in the univariate analysis should be included in the multivariate model. Although this is the main process for the selection of variables, it is not the only one. Variables that reached a borderline association in the univariate analysis (association that did not reach statistical significance, but was close to this) can also be forced into the multivariate model. Investigators should describe the value of the planned level of significance in order to select the univariate results that are worth including in the multivariate model. Most of the time *p* values lower than 0.2 or 0.1 in univariate analyses (an arbitrary assumption that should be described in the study methods) are used.

The third concept necessary to better understand the multivariate model is the concept of confounding variables. The idea of using variables without a clear association in the univariate model can be justified by this concept. An example could be to study the effect of coffee drinking on the development of lung cancer. While these two factors appear to be associated, coffee drinking is also related to smoking, which, in turn, is linked to lung cancer; in this example, coffee drinking is the confounding variable. The fourth concept necessary for the proper interpretation of multivariate models is the need for a minimal number of events of interest per variable. The minimum number of events per variable should be 10–15; if this number is not reached, univariate analysis alone should be performed, dismissing the multivariate model [14, 15]. The fifth concept that can be really useful in clinical research is the inclusion of variables of interest (those addressed in the study hypothesis) that can be forced into the multivariate model regardless of their significance findings in univariate analysis.

The types of multivariate models utilized in clinical research are classified as logistic, linear, or Cox regression, according to the type of dependent variable. In a multivariate model, the difference shows the degree of association of each variable with the presence of the event of interest. In the logistic model, the outcome is assessed by a dichotomous variable, and then continuous variables are usually redistributed on categorical variables, as demonstrated in Table 6.1. In this example, 819 patients who underwent pancreaticoduodenectomy were evaluated according to the presence of complications and variations in the schedules of procedures, as presented in Table 6.3 [1].

Multivariate linear regression can also be used when the dependent variable is continuous, such as the running time in marathons. Another example of the use of multiple linear regressions is the association between tumor markers and continuous clinical variables. Ahn and Ku reported the association of prostate-specific antigen (PSA) level and body mass index (BMI) in young men [16]. After simple linear regressions, as demonstrated in Table 6.4, a multiple linear regression was performed and only BMI, alanine aminotransferase (AST), and creatinine remained related to the PSA level (log10[PSA] = − 0.124[BMI] + 0.063[creatinine] − 0.053[AST] + 0.097; $r = 0.152$, $p < 0.001$) [16].

In time-to-event analysis, Cox regressions include categorical variables to estimate the effect of each variable on the event of interest, according to the time of follow-up. It is important to remember that the time of follow-up is crucial to the internal validity of a study. Studies with short follow-up times underestimate the event rates (death or recurrence, for example), as explained previously.

Table 6.3 Predictors of overall presence of any complications for patients who underwent pancreaticoduodenectomies ($n = 819$) [1]

	Any complications ($n = 405$)				Any leaks or fistulae ($n = 132$)				Death within 90 days ($n = 29$)			
	Univariate		Multivariate		Univariate		Multivariate		Univariate		Multivariate	
Characteristics	*P*-value	OR	95% Cl	*P*-value	*P*-value	OR	95% Cl	*P*-value	*P*-value	OR	95% Cl	*P*-value
Age	0.638	–	–	–	0.474	–	–	–	0.013	1.03	0.98–1.07	0.180
Sex (male versus female)	<0.001	1.48	1.10–1.97	0.009	<0.001	2.00	1.29–3.10	0.002	0.708	–	–	–
BMI (>25 kg/m^2 versus ≤25 kg/m^2)	0.017	1.23	0.91–1.66	0.174	0.024	1.09	0.69–1.71	0.707	0.239	–	–	–
Hypertension	0.485	–	–	–	0.635	–	–	–	0.708	–	–	–
Diabetes mellitus	0.790	–	–	–	0.398	–	–	–	0.474	–	–	–
Cardiac disease	0.025	1.31	0.93–1.84	0.119	0.431	–	–	–	0.003	2.03	0.89–4.58	0.090
Pulmonary disease	0.099	1.45	0.89–2.34	0.134	0.015	1.71	0.95–3.10	0.074	0.055	2.11	0.81–5.50	0.124
Other comorbidities	0.421	–	–	–	0.022	0.51	0.34–0.79	0.002	0.240	–	–	–
ASA class (3 or 4 versus 1 or 2)	0.208	–	–	–	0.028	1.63	1.06–2.50	0.025	0.001	2.09	0.77–5.62	0.144
Greatest diameter of tumour	0.140	–	–	–	0.002	0.83	0.72–0.96	0.013	0.818	–	–	–
Malignant neoplasm (yes versus no)	0.319	–	–	–	0.002	0.44	0.26–0.74	0.002	0.415	–	–	–
Duration of operation	0.005	1.00	0.99–1.00	0.231	0.010	1.00	0.99–1.00	0.492	0.446	–	–	–
Estimated blood loss (100 ml)	0.001	1.03	1.01–1.06	0.017	<0.001	1.03	1.0–1.050	0.084	0.662	–	–	–
Any positive margin	0.109	–	–	–	<0.001	0.13	0.03–0.53	0.005	0.803	–	–	–
Operation month	0.920	–	–	–	0.715	–	–	–	0.736	–	–	–
Operation day (Mon-Wed versus Thurs, Fri)	0.195	–	–	–	0.845	–	–	–	1.000	–	–	
Operation start time												
Before versus after 12.00 h	0.500	–	–	–	0.760	–	–	–	0.690	–	–	–
07.00–11.00 h	0.063	1.03	0.55–1.90	0.937	0.497	–	–	–	0.441	–	–	–
11.01–15.00 h	0.054	0.72	0.38–1.36	0.313	0.422	–	–	–	0.163	–	–	–
After 15.00 h	1.000	–	–	–	0.836	–	–	–	0.219	–	–	–

95% Cl, 95% confidence interval; *ASA* American Society of Anesthesiologists, *BMI* body mass index, *OR* odds ratio

Table 6.4 Linear regression modes according to prostate-specific antigen (PSA) serum levels and clinical parameters

Parameter	Correlation coefficient	*P*-value
Age (yr)	0.002	0.942
Anthropometric measurements		
Height (cm)	0.025	0.289
Weight (kg)	−0.107	<0.001
Body mass index (kg/m^2)	−0.131	<0.001
Hepatic function tests		
Aspartate aminotransferase (U/L)	−0.076	0.001
Alanine aminotransferase (U/L)	−0.079	0.001
Alkaline phosphatase (U/L)	−0.039	0.094
γ-glutamyltransferase (U/L)	−0.046	0.046
Total bilirubin (μmoL/L)	0.038	0.097
Renal function tests		
Urea nitrogen (mmoL/L)	−0.031	0.188
Creatinine (μmoL/L)	0.053	0.024
Creatinine clearance (mL/s)	−0.113	<0.001

Conclusions

In conclusion, we note that Tables, graphs, and curves are crucial for presenting data in a clear fashion. These tools are useful for summarizing and highlighting data, thus enabling readers to gain a better comprehension of items in the text. Tables and graphs must be self-explanatory, and all presented data should be clear enough to show the points addressed without the need for the reader to return to the text. Tables and graphs are usually presented in lectures, and they are also very important tools for making a research article attractive for publication. The various regressions are complementary and should be used as much as the data allows to discharge confounding variables and to identify predictors that might influence the occurrence of the event of interest. Both univariate and multivariate analyses vary according to the types of variables examined, such as binominal, continuous, or time-to-event (with logistic, linear, and Cox regressions, respectively, being used for these three types of variables), and the understanding of these differences is essential for choosing the appropriate analyses and for correctly interpreting the results. Because many issues may be present in regression analyses, it is crucial to have an experienced biostatistician involved in a study's statistical planning and analyses.

References

1. Araujo RL, Karkar AM, Allen PJ, et al. Timing of elective surgery as a perioperative outcome variable: analysis of pancreaticoduodenectomy. HPB (Oxford). 2014;16:250–62.

2. Araujo RL, Gonen M, Herman P. Chemotherapy for patients with colorectal liver metastases who underwent curative resection improves long-term outcomes: systematic review and meta-analysis. Ann Surg Oncol. 2015;22:3070–8.
3. Langer B, Bleiberg H, Labianca R, Shepherd L, Nitti D, Marsoni S, et al. Fluorouracil (FU) Plus L-leucovorin (L-LV) versus observation after potentially curative resection of liver or lung metastases from colorectal cancer (CRC): results of the ENG (EORTC/NCIC CTG/GIVIO) randomized trial. Proc Am Soc Clin Oncol. 2002;21:149a. Abstract 592.
4. Portier G, Elias D, Bouche O, Rougier P, Bosset JF, Saric J, et al. Multicenter randomized trial of adjuvant fluorouracil and folinic acid compared with surgery alone after resection of colorectal liver metastases: FFCD ACHBTH AURC 9002 trial. J Clin Oncol. 2006;24:4976–82.
5. Nordlinger B, Sorbye H, Glimelius B, Poston GJ, Schlag PM, Rougier P, et al. Perioperative chemotherapy with FOLFOX4 and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC Intergroup trial 40983): a randomised controlled trial. Lancet. 2008;371:1007–16.
6. Nordlinger B, Sorbye H, Glimelius B, Poston GJ, Schlag PM, Rougier P, et al. Perioperative FOLFOX4 chemotherapy and sur gery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC 40983): long-term results of a randomised, controlled, phase 3 trial. Lancet Oncol. 2013;14:1208–15.
7. Adam R, Bhangui P, Poston G, Mirza D, Nuzzo G, Barroso E, et al. Is perioperative chemotherapy useful for solitary, metachronous, colorectal liver metastases? Ann Surg. 2010;252:774–87.
8. Reddy SK, Tsung A, Marsh JW, Geller DA. Does neoadjuvant chemotherapy reveal disease precluding surgical treatment of initially resectable colorectal cancer liver metastases? J Surg Oncol. 2012;105:55–9.
9. Ihemelandu C, Levine EA, Aklilu M, Yacoub G, Howerton R, Bolemon B, et al. Optimal timing of systemic therapy in resect able colorectal liver metastases. Am Surg. 2013;79:422–8.
10. Araujo RL, Pantanali CA, Haddad L, et al. Does autologous blood transfusion during liver transplantation for hepatocellular carcinoma increase risk of recurrence? World J Gastrointest Surg. 2016;8:161–8.
11. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. 2012;366:2074–84.
12. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010;19:227–9.
13. Araujo R, Gonen M, Allen P, et al. Comparison between perioperative and postoperative chemotherapy after potentially curative hepatic resection for metastatic colorectal cancer. Ann Surg Oncol. 2013;20:4312–21.
14. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 2001;54:979–85.
15. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373–9.
16. Ahn JO, JH K. Relationship between serum prostate-specific antigen levels and body mass index in healthy younger men. Urology. 2006;68:570–4.

Survival Analysis

7

Benjamin Haaland and Mina Georgieva

7.1 Introduction

Survival endpoints are the outcomes most commonly used to measure the efficacy of new cancer-directed therapy in phase III clinical trials. Therefore, it is crucial for clinicians and investigators to understand the basic concepts of survival analyses and their interpretation. Survival and its derivatives, such as progression-free and disease-free survival, can be measured as categorical variables when subjects are classified as either alive or deceased at a specific time point; in this case the outcome variable is binomial and is expressed as the survival rate at a given time. However, most frequently, survival is measured as a time-to-event variable. These variables are continuous measures that do not encompass all information, but rather include missing data. This is because some data are unknown or incomplete at the time of analysis. For example, overall survival is calculated from a pre-defined date, e.g., the first day of treatment, until death, as a continuous numerical value (weeks, for instance). However, in real life, many patients are lost to follow-up, remain alive at the time of study termination, or may simply withdraw from trial participation before the study end. Actually it is very unlikely that investigators have all the data about the date of death of all study subjects because it may take a long period of time until all deaths occur and/or because it is impossible to avoid any loss of information about patients' deaths. Because we cannot simply exclude subjects who have not reached the event of death (or progression, if the outcome variable is progression-free survival, for example), some individuals are classified as

B. Haaland, Ph.D. (✉)
University of Utah, Salt Lake City, UT, USA
e-mail: ben.haaland@hci.utah.edu

M. Georgieva, M.S.
Georgia Institute of Technology, Atlanta, GA, USA

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_7

'censored' in the survival analysis. The time a subject remains on study, regardless of achieving the event or being censored, is included in the survival analysis.

This chapter discusses approaches and analytic techniques related to time-to-event endpoints, which are common in cancer studies. Ordinarily, interpretation and comparison of time-to-event data are made complex by the presence of right-censoring, where it is only known that an event has not occurred up to some time. Care is required to correctly interpret summaries and comparisons for this type of data, as discussed below.

Note:

This chapter includes several technical sections, provided in text boxes as shown in Technical Material 000 below, which are included for completeness and as a resource for quantitatively oriented researchers. For readers who are interested in a high-level overview, these sections may safely be skipped. For readers interested in a more comprehensive introduction to survival analysis, a reasonable starting point could be *Survival analysis: a self-learning text* [1]. All analyses and examples in this chapter use the statistical software R [2]. R is a very powerful and freely available software package that can be downloaded at https://www.r-project.org/. For readers interested in using R to analyze their own data, a useful resource could be *Applied Survival Analysis using R* [3].

Technical Material 000: Technical material text box
Technical material

7.2 Time-to-Event and Right-Censored Data

7.2.1 Time-to-Event Data

Endpoints, or outcomes of interest, in many cancer studies are times-to-event. For example, time to death or time to progression or death. Consider as an example the CRYSTAL study comparing cetuximab with FOLFIRI to FOLFIRI alone as first-line therapy for advanced or metastatic colorectal cancer [4]. The study's primary endpoint is time to progression or death, i.e., progression-free survival (PFS), while a secondary endpoint is time to death, i.e., overall survival (OS). For comparison of first-line therapies in a setting with post-progression therapy, OS comparisons can be contaminated by unknown post-progression treatments.

Time-to-event data, especially in a cancer context, are commonly right-skewed with a non-negligible probability of much larger than typical times-to-event. These longer than typical times-to-event data represent patients with long-term survival. However, given the time and cost constraints of long-term follow-up, as well as staggered enrollment times leading to different follow-up times, many times-to-event are only partially observed. For example, we may only observe that the event of interest did not happen up to a particular time.

Typical goals that a researcher may have with time-to-event data are the summarizing of survival distributions (for example, median survival times or proportions of more than 5-year survivors), comparing survival distributions, and modeling the impact of predictors on the survival distribution.

7.2.2 Right-Censoring

Censoring occurs when a variable is only partially observed. In particular, right-censoring occurs when it is only observed that the variable's true value is larger than (to the right of) a particular value. For example, suppose the endpoint of interest is survival. A particular patient has undergone 20 months of follow-up at the time of analysis, and is still alive. Then s/he is (right-) censored at 20 months. A common notation for this observed time is 20+, indicating that the true survival time exceeds 20 months.

Some of the common reasons for right-censoring in cancer studies include administrative censoring, loss to follow-up, and competing risks. Administrative censoring occurs when a study observation time ends, say for either analysis or because of study closing. Administrative censoring can occur owing to a fixed duration of follow-up for each patient or owing to a fixed time of study (or data) closing. Importantly, the censoring mechanism in administrative censoring is statistically independent of the event times. Loss to follow-up, on the other hand, occurs when a subject exits the study, for whatever reason, and their endpoint cannot be measured after their last follow-up time. A time-to-event measurement can also be censored by the occurrence of a competing risk. For example, progression time cannot be measured in a patient who dies before they progress.

Specialized techniques have been developed for estimation and regression in the context of right-censored data, but almost all of these techniques make the assumption of, and depend strongly on, independence between the event time and the censoring mechanisms. Administrative censoring schemes generally meet this assumption. Independence between event times and censoring is more questionable for censoring that is driven by loss to follow-up and competing risks. For example, a patient who is doing poorly may be both more likely to withdraw from the study early and more likely to have negative events early. Similarly, a patient who is more likely to progress early may be more likely to die early. Consider time-to-progression (TTP) in comparison to PFS. In TTP, the event of interest is progression, subject to right-censoring owing to competing risks such as death and loss to follow-up, as well as administrative censoring. In PFS, on the other hand, the event of interest is a composite of progression or death, subject to right-censoring owing to loss to follow-up, as well as administrative censoring. Because death times might be expected to be positively associated with progression times in most situations of studies of advanced cancer (that is, the assumption of independence between event times and censoring time is not met for TTP), regulatory agencies generally prefer PFS over TTP.

7.2.3 Example

Consider a toy example of a randomized clinical trial (RCT), where patients enroll over a 24-month period and are randomly assigned to arm A or arm B. The trial closes at 27 months, so patients who enrolled early in the trial have as much as 27 months of follow-up, while patients who enrolled late in the trial have as little as 3 months of follow-up. The upper panels of Fig. 7.1 show patient enrollment and event or censoring times *over the course of the study*, while the lower panels show patient follow-up times. For both panels, a line extends from either the patient's enrollment time or start of follow-up time (time 0) to either their event time (say PFS) or their time from start of follow-up to the event, subject to right-censoring. Lines that terminate in a filled-in dot indicate that the event actually occurred at that time, while lines that terminate in an empty dot indicate that a censoring occurred at that time, and in turn the event of interest occurred in the future. The upper panels have been sorted according to enrollment times and the lower panels have been sorted according to follow-up times, for ease of viewing. Notice that in the lower panels, the PFS times for arm B appear to have a tendency to be larger than those for arm A. Quality estimation of the distribution of survival times and testing for differences between survival distributions, using right-censored data, will be discussed in the next two sections.

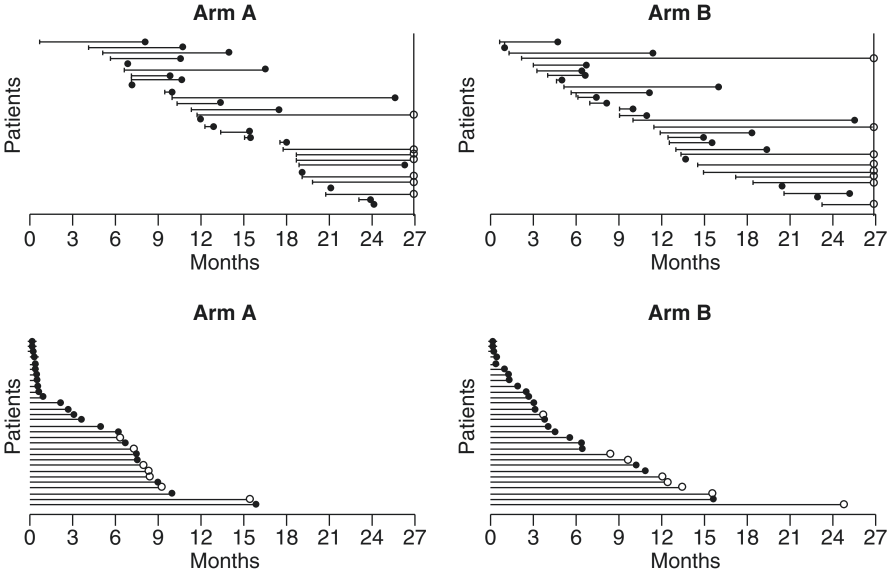


Fig. 7.1 Patient enrollment and event or censoring times over the course of the study, *shown in* the *upper panels*; patient follow-up times are *shown in* the *lower panels*

7.3 Survival Curve Estimates

7.3.1 Survival Distribution

In this section, we discuss estimation of the survival distribution (i.e., the distribution of event times), as well as related quantities, such as median survival times and the probability of long-term survival, using right-censored time-to-event data. The central quantity for each of these is commonly the Kaplan-Meier (KM) estimator of the survival curve. Parametric survival curve estimates will be discussed after the KM estimator. As we will see below, the KM estimator involves a few slightly convoluted calculations, and a very natural question is: why not use something simpler? The answer is that the KM estimator makes full use of the data, including the censored observations, while not estimating quantities that require assumptions beyond independence between the event and censoring times. For example, unless one observes event times (not censorings) in the tail, or right-hand side, of the distribution, then assumptions are needed to estimate any quantity that depends on the tail of the distribution, such as the mean time-to-event. In general, the strong right-skew present in many survival distributions makes the mean an inappropriate measure of the central tendency for most time-to-event data, even when the right tail can be estimated. That is why the median, rather than the mean, is the measure of central tendency used to report time-to-event variables.

As an example, consider using the empirical cumulative distribution function (ECDF) to estimate the survival distribution. Notably, if one has a quality estimate of the survival distribution, then an estimate of the median can be taken as a/the point at which there is equal probability of survival times being longer and shorter. Similarly, an estimate of say, the 2-year survival rate, is simply *one minus* an evaluation of the cumulative distribution function at 2 years. The ECDF estimates the probability that a *random* survival time T is less than or equal to a particular time of interest t, $P(T \leq t)$, as the observed proportion of event times that are less than or equal to t, $\hat{P}(T \leq t) = \frac{1}{n}\#(T_i \leq t)$, where the notation $\#(A)$ denotes the number of observations, T_i's, in event A. One way to misuse the ECDF would be to simply treat the censoring times as event times, but this would clearly bias the survival distribution estimate to the left. The ECDF would indicate that survival times tended to be shorter than they actually were. A step towards a more reasonable survival distribution estimate would be to simply omit the censored observations, then use the ECDF distribution estimate on the uncensored observations. This approach is problematic for at least two reasons. First, the resulting estimator is highly inefficient if the proportion of censored observations is high, as it is in many cancer studies. Second, the quantity that is being estimated is the distribution *conditional on not being censored*. For example, if every patient in the study were followed-up for exactly 1 year, then the ECDF distribution would estimate the distribution of survival times conditional on the survival times being less than 1 year. The KM estimator addresses these censoring-related issues and in fact is equivalent to the ECDF when there is no censoring.

7.3.2 Kaplan-Meier Survival Curve Estimator

In clinical research, the KM estimator is a non-parametric estimator of the proportion of patients living for a period of time after receiving a therapeutic intervention. It is one of the most frequently used methods to estimate survival times in general medical research; in oncology research, KM survival curve estimates are mostly used to report probabilities of death, progression, or recurrence.

The KM estimator targets the so-called *survival function* $S(t) = P(T > t)$ (notice that this equals one minus the cumulative distribution function, $P(T \leq t)$) and leverages a key identity from conditional probability. Recall the *definition* of conditional probability, $P(A|B) = P(A, B)/P(B)$, which can be rearranged to express a joint probability as the product of a conditional and a marginal probability, $P(A, B) = P(A|B) \times P(B)$. The KM estimator uses this fact recursively to estimate the survival distribution cleanly in the presence of right-censoring, as shown in Technical Material 111.

Technical Material 111: Construction of the Kaplan-Meier survival curve estimator

Suppose that the particular t of interest is less than or equal to the longest follow-up time, whether follow-up was ended by an event or censoring, and write the unique follow-up times as $t_1 < t_2 < \cdots < t_k$. Then, if the time of interest t is between the observed times t_k and t_{k+1}, $t_k < t \leq t_{k+1}$,

$$P(T > t) = P(T > t, T > t_k, \ldots, T > t_1),$$

since $T > t$ if and only if $T > t$ and $T > t_k$ and $T > t_{k-1}$ and so on. This joint probability can, in turn, be expressed as the product of a conditional and marginal probability,

$$\begin{aligned} P(T > t) &= P(T > t | T > t_k, \ldots, T > t_1) \times P(T > t_k, \ldots, T > t_1) \\ &= P(T > t | T > t_k) \times P(T > t_k, \ldots, T > t_1), \end{aligned}$$

where the second equality follows from the fact that $T > t_k$ if and only if $T > t_k$ and $T > t_{k-1}$ and $T > t_{k-2}$ and so on. Applying this argument next to $P(T > t_k, \ldots, T > t_1)$ and repeating allows us to rewrite $P(T > t)$ as a product of conditional probabilities and one marginal probability,

$$P(T > t) = P(T > t | T > t_k) \times P(T > t_k | T > t_{k-1}) \times \cdots \times P(T > t_2 | T > t_1) P(T > t_1).$$

Each of these component probabilities can be estimated cleanly in the presence of right-censoring. First, consider the conditional probabilities $P(T > t_k|T > t_{k-1})$. These can be rewritten in terms of their complements, since $P(\text{not } A) = 1 - P(A)$. Let d_k denote the number of events (not censorings) that occur at time t_k and let n_k denote the number of subjects *at risk* (neither censored nor dead) just *before* time t_k. Recall that no events or censorings occur between t_{k-1} and t_k, and each of the at-risk units is independent. This means that conditional on n_k, the number of events d_k has a binomial distribution with n_k trials and underlying probability $P(T \le t_k|T > t_{k-1})$, implying a maximum likelihood estimate (parameter setting which makes observed data most typical) of $P(T > t_k|T > t_{k-1})$,

$$\hat{P}\left(T > t_k \middle| T > t_{k-1}\right) = 1 - \hat{P}\left(T \le t_k \middle| T > t_{k-1}\right) = 1 - \frac{d_k}{n_k}.$$

A similar argument indicates that $\hat{P}\left(T > t \middle| T > t_k\right) = 1 - \frac{0}{n_k} = 1$ and $\hat{P}\left(T > t_1\right) = 1 - \frac{d_1}{n_1}$.

Putting the above development together gives the so-called KM survival estimator.

$$\hat{S}(t) = \hat{P}(T > t) = \prod_{k:t_k \le t} \left(1 - \frac{d_k}{n_k}\right),$$

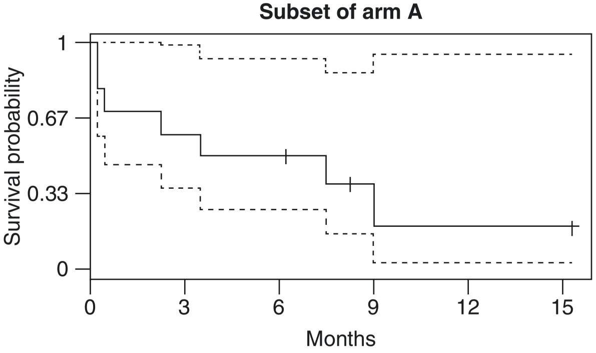
where the notation $\Pi_{k:t_k \le t}$ denotes the *product* of terms with $t_k \le t$, d_k denotes the number of events (not censorings) that occur at time t_k, and n_k denotes the number of subjects *at risk* (neither censored nor dead) just *before* time t_k [5]. Notice that the KM survival estimator decreases only at event times (since times with only censorings have $d_k = 0$). Visually, the KM estimate has a "staircase" pattern, with drops at observed event times. In the event that the last follow-up time is a censoring (so that the KM does not go all the way to zero), the KM curve is usually only extended to this last follow-up time, reflecting the fact that we simply do not have observations providing information about the right tail of the distribution.

7.3.3 Example

Consider the subset of ten observations drawn from the arm A data described in the example above and given in Table 7.1. Times annotated with a "+" indicate right-censored observations. We demonstrate calculation of the KM estimate in columns 1 through 6 of Table 7.1. In the first column, unique follow-up times (t_k) are listed. Notice we have two 0.25-month follow-up times. In the second, third, and fourth columns, we list number at risk (n_k), number of events (d_k), and number of

Table 7.1 Subset of ten observations drawn from arm A data[a], shown in the upper left panel. The Kaplan-Meier survival curve estimate is shown in upper right panel

Months: 0.25, 0.25, 0.50, 2.25, 3.50, 6.25+, 7.50, 8.25+, 9.00, 15.25+



Unique times (t_k)	Number at risk (n_k)	Number of events (d_k)	Number of censorings	$1-\frac{d_k}{n_k}$	$\hat{S}(t)$	$\frac{d_k}{n_k(n_k-d_k)}$	$\hat{\mathrm{V}}\mathrm{ar}\left(\log \hat{S}(t)\right)$	(lower, upper)
0.25	10	2	0	0.800	0.800	0.025	0.025	(0.587, 1.000)
0.50	8	1	0	0.875	0.700	0.018	0.043	(0.467, 1.000)
2.25	7	1	0	0.857	0.600	0.024	0.067	(0.362, 0.995)
3.50	6	1	0	0.833	0.500	0.033	0.100	(0.269, 0.929)
6.25	5	0	1	1.000	0.500	0.000	0.100	(0.269, 0.929)
7.50	4	1	0	0.750	0.375	0.083	0.183	(0.162, 0.868)
8.25	3	0	1	1.000	0.375	0.000	0.183	(0.162, 0.868)
9.00	2	1	0	0.500	0.188	0.500	0.683	(0.037, 0.948)
15.25	1	0	1	1.000	0.188	0.000	0.683	(0.037, 0.948)

Step-by-step calculation of Kaplan-Meier and corresponding 95% confidence interval are shown in the lower cells

[a] See text for description of the study for which arm A data are shown and for definitions of all terms

censorings. At each follow-up time, we compute the factor $1-\frac{d_k}{n_k}$ in column 5, which is the *factor* by which the KM drops at each follow-up time. Notice that the KM drops only if there is an observed event. Mechanically, $\hat{S}(t)$ equals 1 until the first event time, then the new value of $\hat{S}(t)$ is calculated by multiplying its previous value by the factor $1-\frac{d_k}{n_k}$. Notice that until censoring times occur, $\hat{S}(t)$ is simply equal to one minus the ECDF, or the proportion of event times exceeding t. If no censoring occurs, then $\hat{S}(t)$ equals one minus the ECDF for all t. In the upper right entry in Table 7.1, we see a plot of the KM curve (solid line), with censoring times marked with a vertical tick. The next two columns in Table 7.1 are used to compute pointwise 95% confidence intervals, as given in the final column, and to conduct hypothesis tests.

Several estimates can be *read* directly from the KM survival curve estimate. For example, an estimate of the probability that a patient lives beyond 12 months is $\hat{P}(T>12)=\hat{S}(12)=0.188$. By construction, the KM curve is right-continuous, meaning that at one of the staircase drops, the KM curve actually equals the value of the lower step. For example, an estimate of the probability that a patient lives beyond 7.5 months is $\hat{P}(T>7.5)=\hat{S}(7.5)=0.375$. From the other perspective, we can *read* quantiles from the KM estimate by plugging in y-coordinates to find the corresponding x-coordinates. Recall that, *roughly speaking*, a quantile of the distribution of T is a particular time t_* so that there is some desired probability of $T \le t_*$. In fact, there are a number of reasonable ways one might apply the above *rough* definition when faced with actual data. For example, suppose we would like to estimate the median survival time, the t_* that makes $P(T \le t_*) = 0.5$. Referring to the KM estimate in the upper right panel of Table 7.1, we see that the horizontal line $y = 0.5$ intersects with the survival curve estimate from $3.5 \le t < 7.5$. A typical choice is to take the midpoint of this interval $t_* = 5.5$ as the median estimate. Suppose, on the other hand, one wishes to estimate the 75th percentile, the t_* that makes $P(T \le t_*) = 0.75$. In terms of the survival function this t_* has $1 - S(t_*) = 0.75$ or $S(t_*) = 1 - 0.75 = 0.25$. Now, the horizontal line $y = 0.25$ intersects the KM curve at the "drop" at $t = 9.0$. From another perspective, the horizontal line $y = 0.25$ does not intersect the KM curve at all; there is no t_* satisfying the above relation. Common practice, however, is to take $t_* = 9.0$ as the estimate of the 75th percentile. Notably, since the final follow-up time is a right-censoring at 15.25 months, the survival curve estimate *stops* at 15.25. This makes the mean, which is commonly used to summarize the location of data that is not right-censored, *impossible* to estimate. To see this, consider two extreme scenarios. First, suppose all remaining events occur at time 16, then the mean is something less than 16 months. Second, suppose all remaining subjects do not have events (event times equal infinity), then the mean equals infinity. We have no information to distinguish between these extremes, or the more likely middle ground.

7.3.4 Inference

If we want to perform *inference* on our survival curve estimate, that is, to statistically quantify the uncertainty intrinsic to the estimator, we need the sampling distribution of $\hat{S}(t)$. In fact, it can be shown that as more and more data accumulate, the sampling distribution of $\hat{S}(t)$, as well as $\log \hat{S}(t)$, is more and more closely approximated by a normal distribution centered at the true/target value, $S(t)$ or $\log S(t)$, and whose variance can be estimated using the techniques outlined in Technical Material 222. Throughout this section the logarithm refers to the *natural* (base $e \approx 2.718$) logarithm.

Technical Material 222: Construction of hypothesis tests and confidence intervals for the survival curve

It can be shown that a quality estimate of the variance of the Kaplan-Meier (KM) survival estimator at time t is given by Greenwood's formula,

$$\widehat{\text{Var}}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{k:t_k \le t} \frac{d_k}{n_k\,(n_k - d_k)}$$

where the notation $\Sigma_{k:t_k \le t}$ denotes the summation of terms with $t_k \le t$ [6]. Then, for small $\widehat{\text{Var}}\left(\hat{S}(t)\right)$, the KM survival estimate is approximately normally distributed with mean $S(t) = P(T > t)$, the true (target) survival function at time t, and variance $\widehat{\text{Var}}\left(\hat{S}(t)\right)$,

$$\hat{S}(t) \sim \mathcal{N}\left(S(t), \widehat{\text{Var}}\left(\hat{S}(t)\right)\right).$$

More commonly, however, inference is performed on the (natural) logarithm scale. Similar to the untransformed KM, $\log \hat{S}(t)$ is also approximately normally distributed with mean $\log S(t)$ and variance given by

$$\widehat{\text{Var}}\left(\log \hat{S}(t)\right) = \sum_{k:t_k \le t} \frac{d_k}{n_k\,(n_k - d_k)}$$

nearly identical to the untransformed case, but without the $\hat{S}(t)^2$ term. Once again, for small $\widehat{\text{Var}}\left(\log \hat{S}(t)\right)$,

$$\log \hat{S}(t) \sim \mathcal{N}\left(\log S(t), \widehat{\text{Var}}\left(\log \hat{S}(t)\right)\right).$$

This large sample distribution can be used to build a confidence interval for the true survival curve $S(t)$

$$\exp\left\{\log \hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\log \hat{S}(t)\right)}\right\},$$

where exp{x} denotes the natural exponential e^x and $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the standard normal distribution; for example, if $\alpha = 0.05$, then $z_{1-\alpha/2} = z_{0.975} = 1.96$. While the confidence interval is positive by construction, it can be larger than 1, and is typically truncated at 1 should the computed upper limit exceed 1 (since we are estimating a probability, we know it must be between 0 and 1). Hypothesis tests (at level α) for $\mathrm{H}_0 : S(t) = \theta_0$ can be conducted by checking whether θ_0 is contained in the confidence interval. Note that the confidence interval above is *pointwise* in the sense that it holds for any particular time of interest t, but not all times of interest simultaneously. In particular, a confidence *envelope* that has a $1 - \alpha$ probability (before data collection) of containing the *entire* survival curve would need to be a bit wider.

7.3.5 Example

The confidence intervals described above can be used to annotate a plot of the KM survival curve estimate. The confidence intervals also have a “staircase” pattern, with drops at observed event times. The final column of Table 7.1 gives 95% *log-scale* confidence intervals (truncated above at 1) as per the equation above in Technical Material 222. The upper right panel of Table 7.1 shows the confidence intervals as dashed lines. Now, we can “read” confidence intervals and tests from the plot by examining where vertical or horizontal lines representing the time or quantile of interest intersect the confidence intervals. For example, 95% confidence intervals on $P(T > 12)$ and $P(T > 7.5)$ are respectively given by (0.037, 0.948) and (0.162, 0.868). Null hypothesized values falling within the intervals (including endpoints typically) are not rejected, while null hypothesized values not falling within the intervals are rejected. Similarly, respective 95% confidence intervals on the median and 75th percentile are given by (0.50, not reached) and (7.50, not reached). Notice that the upper confidence boundary does not reach either the median or the 75th percentile.

7.3.6 Nelson-Aalen Estimator

We saw above that the unconditional probability of survival as a function of time can be described using the survival function. Note that as a subject ages, the conditional probability of survival is dependent on the length of time the subject has not experienced an event. Sometimes, capturing the process of aging and the underlying mechanisms driving events is of more interest than estimating the unconditional survival distribution. The quantity that captures the process of aging is the hazard function, or force of mortality. Construction of the common Nelson-Aalen hazard function estimator, as well as related inference, is provided in Technical Material 333.

Notably, the Nelson-Aalen estimate of the survival function, $\hat{S}(t) = e^{-\hat{\Lambda}(t)}$, and the K-M survival estimate are asymptotically equivalent. On the other hand, in small samples, there is evidence that Nelson-Aalen is better for estimating the cumulative hazard function (see below), while KM is better for estimating the survival function.

Technical Material 333:Construction and inference for Nelson-Aalen hazard function estimator

The hazard function, or *force of mortality*, has a natural interpretation as the event rate in the next small time segment conditional on the subject having survived up to a particular time of interest *t*,

$$\frac{P(t < T \le t + \Delta t / T > t)}{\Delta t} = \frac{P(t < T \le t + \Delta t)}{\Delta t \times P(T > t)} = \frac{P(t < T \le t + \Delta t)}{\Delta t \times S(t)}$$

where Δt is taken to be a small increment in time. It can be shown that as Δt approaches zero, this quantity converges to

$\lambda(t) = f(t)/S(t)$,

where $f(\cdot)$ denotes the probability density function of the event time distribution. Notice that $\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log S(t)$ (chain rule for differentiation). Now, let $\Lambda(t)$ denote the cumulative hazard function that captures the accumulation of hazard over time. Putting these together establishes a useful link between the survival function $S(t)$ and the cumulative hazard function $\Lambda(t)$:

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u = -\int_0^t \frac{\mathrm{d}}{\mathrm{d}u}\log S(t)\,\mathrm{d}u = -\log S(t).$$

Although the cumulative hazard function $\Lambda(t)$ lacks an intuitive interpretation, it offers an alternative characterization of the survival function in terms of the identity

$S(t) = e^{-\Lambda(t)}$.

Now, suppose the particular time of interest *t* is less than or equal to the longest follow-up time. The Nelson-Aalen estimator is a direct estimator of $\Lambda(t)$ that is based on a maximum likelihood estimate of the hazard rates, and is given by

$$\hat{\Lambda}(t) = \sum_{k:t_k \le t} \frac{d_k}{n_k}$$

Where, once again, $t_1 < t_2 < \cdots < t_k$ are the unique follow-up times, d_k denotes the number of events (not censorings) that occur at time t_k, and n_k denotes the number of subjects *at risk* (neither censored nor dead) just *before* time t_k. The Nelson-Aalen estimator is an increasing step function with increments of size $\frac{d_k}{n_k}$ occurring only at event times. The ratio $\frac{d_k}{n_k}$ quantifies the hazard at each unique event time, and the sum captures the accumulation of hazard over time.

Similarly to the KM survival estimate, as more and more data points accumulate, the sampling distribution of $\hat{\Lambda}(t)$ can be approximated by a normal distribution. Further, an estimate of the variance of the Nelson-Aalen estimator at time t is given by

$$\widehat{\text{Var}}\left(\hat{\Lambda}(t)\right) = \sum_{k:t_k \le t} \frac{(n_k - d_k)\, d_k}{(n_k - 1)\, n_k^{\,2}}$$

The Nelson-Aalen survival estimate is approximately normally distributed, with mean, $\Lambda(t)$, the true (target) cumulative hazard function, and variance approximately $\widehat{\text{Var}}\left(\hat{\Lambda}(t)\right)$. The Nelson-Aalen estimator can be used to build an estimate and confidence interval for the true survival curve $S(t)$ using the identity $S(t) = e^{-\Lambda(t)}$, and this estimator is asymptotically equivalent to the KM estimator.

7.3.7 Example

Consider the subset of ten observations drawn from the arm A data described in the example above. In the left panel of Fig. 7.2, we see a plot of the Nelson-Aalen cumulative hazard curve (solid lines). In the right panel of Fig. 7.2, we see a plot comparing the KM survival estimate with the Nelson-Aalen estimate of the survival function.

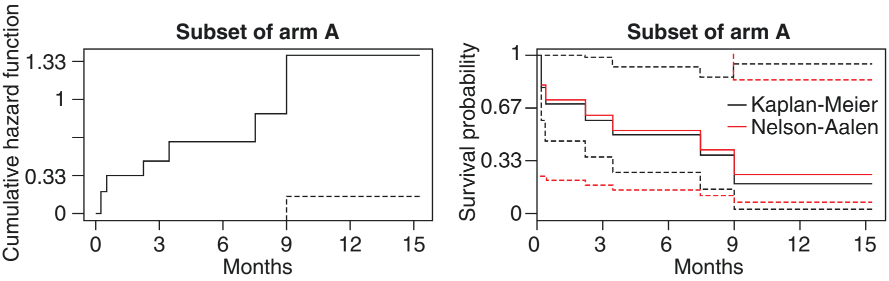


Fig. 7.2 Nelson-Aalen estimate of cumulative hazard *shown in the left panel*. Nelson-Aalen and Kaplan-Meier survival curve estimates are compared *in the right panel*

7.3.8 Parametric Survival Models

We saw above two non-parametric methods, the KM estimate of the survival function and the Nelson-Aalen estimate of the cumulative hazard function. Parametric survival models offer a different approach to survival analysis in which all but a few parameters in the model are explicitly specified, including the survival and hazard functions. These models are based on assumptions about the form of the underlying distribution of the survival time. Although parametric models require additional assumptions, they allow easier estimation, extrapolation to the tail of the survival curve, and more complex analysis. Typically, the unknown parameters are estimated via maximum likelihood (values of parameters that make data most typical). The underlying distribution for the survival time can be chosen based on the shape of the hazard or survival function. Once the distribution is specified (and its corresponding probability density function $f(t)$ is expressed in terms of the unknown parameters), the survival function and the hazard function can be found using the formulas discussed in Technical Material 333. A brief technical example is described in Technical Material 444.

Technical Material 444: Brief exponential model technical example

Consider, as an example, the simplest parametric survival model—the exponential model with unknown event rate μ. The probability density function is given by $f(t) = \mu e^{-\mu t}$. The survival function $S(t)$ can be obtained by integrating $f(\cdot)$ from time t to infinity, $S(t) = P(T > t) = \int_t^\infty f(u)\,du = \int_t^\infty \mu e^{-\mu t}\,\mathrm{d}u = e^{-\mu t}$.

Definitions of the hazard function and the cumulative hazard function yield $\lambda(t) = \frac{f(t)}{S(t)} = \mu$ and $\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u = \int_0^t \mu\,\mathrm{d}u = \mu t$, the the respective hazard and cumulative hazard functions for the exponential distribution.

Some of the most commonly used parametric models are the exponential model, the Weibull model, the log-logistic model, and the log-normal model. The exponential has a constant hazard, while the Weibull has either an increasing, decreasing, or constant hazard, depending on its parameters. Both the log-logistic and log-normal models have increasing, then decreasing hazard functions. Generally, the choice to use a parametric model, as well as the choice of which parametric model, is based on graphical summaries of model adequacy, commonly in terms of the cumulative hazard (within groups of similar patients). Table 7.2 shows the survival, hazard, and cumulative hazard functions of these distributions.

7.3.9 Example

Consider the subset of ten observations drawn from the arm A data described in the previous examples. In the left panel in Fig. 7.3, we see the predicted survival probabilities for the exponential model, the Weibull model, the log-logistic model,

Table 7.2 Survival, hazard, and cumulative hazard functions for some of the most common parametric survival models

Distribution	Parameter	$S(t)$	$\lambda(t)$	$\Lambda(t)$
Exponential	$\mu > 0$	$e^{-\mu t}$	μ	μt
Weibull	$\alpha > 0$ shape $\mu > 0$ rate	$e^{-(\mu t)^{\alpha}}$	$\alpha\mu^{\alpha}t^{\alpha-1}$	$(\mu t)^{\alpha}$
Log-logistic	$\alpha > 0$ shape $\mu > 0$ rate	$\frac{1}{1+(\mu t)^{\alpha}}$	$\frac{\alpha\mu^{\alpha}t^{\alpha-1}}{1+(\mu t)^{\alpha}}$	$\log[1+(\mu t)^{\alpha}]$
Log-normal[a]	μ $\sigma > 0$ scale	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{\phi\left(\frac{\log t-\mu}{\sigma}\right)}{\sigma t\left[1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right]}$	$\frac{\phi\left(\frac{\log t-\mu}{\sigma}\right)}{\sigma t}$

See text for definitions of all terms

[a] *ϕ and* Φ are the standard normal distribution probability density function and cumulative distribution function, respectively

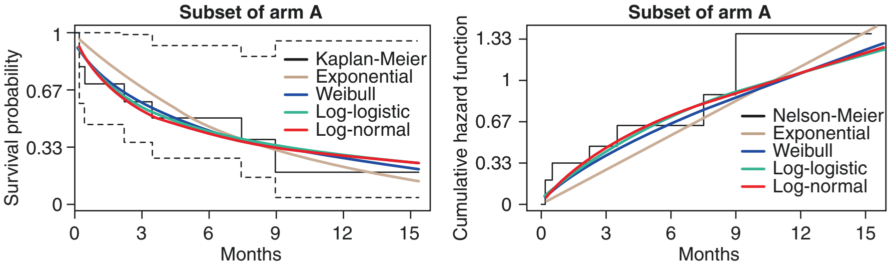


Fig. 7.3 Comparison of parametric survival models and Kaplan-Meier model, *shown in the left panel*. Comparison of parametric cumulative hazard models and Nelson-Aalen is *shown in the right panel*

and the log-normal model, each with parameters estimated via maximum likelihood, along with the KM estimator. In the right panel in Fig. 7.3, we see the corresponding estimated cumulative hazard function, along with the Nelson-Aalen estimate.

7.4 Cox Proportional Hazards Model

We saw above the techniques for estimating quantities relating to a single survival curve, not depending on covariates. In many situations, we would like to test whether group membership, or a more general set of covariates, impacts the survival distribution, or more generally, we would like to estimate the impact of group membership or covariates. Most typically these types of tests and estimates are generated in the context of the so-called *Cox proportional hazards* regression model, which is a quite general regression modeling framework for right-censored data. However, we consider the simple case where there are two groups to be compared, and the associated *log-rank test*, to develop an intuitive understanding of the overall modeling approach.

7.4.1 Log-Rank Test for Comparing Two Groups

The essential idea behind the log-rank test (and the Cox proportional hazards model) is to *condition* on the numbers at risk in each group and the *total* number of events at each unique event time. This *conditioning* means that for each event time, we consider the total number of events and numbers at risk in each group as fixed or given, and the randomness arises from which group, or subjects, these events occur for. Construction of the log-rank test for comparing two groups is detailed briefly in Technical Material 555.

Technical Material 555: Construction of log-rank test

Consider a particular unique event time t_k, numbers at risk in groups 1 and 2 of n_{k1} and n_{k2}, and total number of events d_k. We have d_k events and $n_k - d_k$ non-events, where $n_k = n_{k1} + n_{k2}$. Under the null hypothesis that the survival distributions in the two groups are equal, the number of events occurring in group 1 d_{k1} (or equivalently group 2) at time t_k represents a random number of events present in a sample of size n_{k1} drawn from the full-size n_k sample. In particular,

$$P\left(D_{k1} = d_{k1}\right) = \binom{d_k}{d_{k1}}\binom{n_k - d_k}{n_{k1} - d_{k1}} / \binom{n_k}{n_{k1}},$$

where the notation $\binom{a}{b} = a!/\left(b!(a-b)!\right)$ denotes the number of ways to choose b objects from a objects for integers $a \geq b$. This distribution is a (central) hypergeometric distribution with mean $\mu_{k1} = d_k n_{k1}/n_k$ and variance

$$V_k = \frac{d_k\left(n_k - d_k\right)n_{k1}n_{k2}}{n_k^2\left(n_k - 1\right)}.$$

It can be shown that, under the null hypothesis that the survival distributions in the two groups are the same, the test statistic $Z = \sum_k \left(d_{k1} - \mu_{k1}\right) / \sqrt{\sum_k V_k}$ is distributed as approximately standard normal $N(0,1)$.

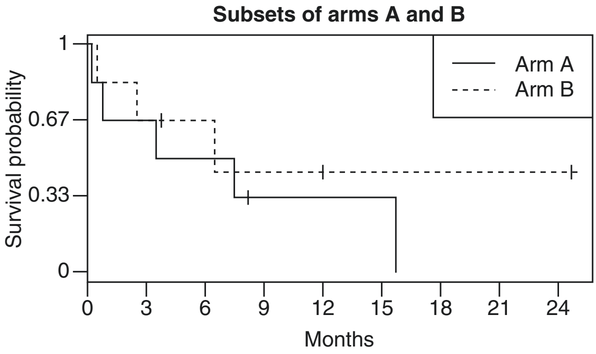
The so-called *log-rank statistic* is typically Z^2 (in Technical Material 555), which is approximately distributed as chi-squared with 1 degree of freedom χ_1^2 under the null. Notably, if one survival curve tends to drop more quickly (or more slowly) than the other overall, then this statistic will be larger than a typical χ_1^2 random deviation, and the corresponding *p*-value will be small. On the other hand, if the two survival distributions differ but one does not consistently drop more or less quickly than the other, then the log-rank statistic may not be large relative to a typical χ_1^2 random draw. The (standard) log-rank test does not have high power for detecting differences between survival distributions when one is not consistently dropping more quickly than the other. When differences between the rates of decline of the survival functions are expected to be concentrated in a particular time segment, weighted versions of the log-rank test, focusing on the time segment of interest, are available, and have improved power.

7.4.2 Example

Consider small subsets of six observations from each of arms A and B from the example described above and shown in the upper left panel of Table 7.3. Respective KM survival curves are shown in the upper right panel of Table 7.3, arm A in solid black and arm B in dashed black. The columns in the lower portion of Table 7.3

Table 7.3 Subsets of six observations drawn from each of arms A and B[a], shown in the upper left panel. Kaplan-Meier survival curve estimates are shown in the upper right panel

Subset of Arm A
Months: 0.25, 0.75, 3.50, 7.50, 8.25+, 15.75



Subset of Arm B
Months: 0.50, 2.50, 3.75+, 6.50, 12.00+, 24.75+

t_k	n_{k1}	n_{k2}	d_{k1}	d_{k2}	μ_{k1}	$d_{k1} - \mu_{k1}$	V_k
0.25	6	6	1	0	0.500	0.500	0.250
0.50	5	6	0	1	0.455	−0.455	0.248
0.75	5	5	1	0	0.500	0.500	0.250
2.50	4	5	0	1	0.444	−0.444	0.247
3.50	4	4	1	0	0.500	0.500	0.250
3.75	3	4	0	0	–	–	–
6.50	3	3	0	1	0.500	−0.500	0.25
7.50	3	2	1	0	0.600	0.400	0.240
8.25	2	2	0	0	–	–	–
12.00	1	2	0	0	–	–	–
15.75	1	1	1	0	0.500	0.500	0.250
24.75	0	1	0	0	–	–	–

Step-by-step calculation of log-rank test components is shown in the lower cells

[a] See text for description of the study for which arm A and arm B data are shown and for definitions of all terms

show the observed times t_k, numbers at risk in arms A and B n_{k1} and n_{k2}, and the number of events in arms A and B d_{k1} and d_{k2}, as well as the expected number of events in arm A μ_{k1} (under the null hypothesis that the event rates in the two groups are equal), the observed minus the expected event rates $d_{k1} - \mu_{k1}$, and the variances V_k, at each observation time. Summing the observed minus the expected event rates and variancesover *event times*, gives $\sum_k d_{k1} - \mu_{k1} = 1.001$ and $\sum_k V_k = 1.985$. In turn, $Z = \sum_k (d_{k1} - \mu_{k1}) / \sqrt{\sum_k V_k} = 0.710$ and $Z^2 = 0.505$. We can compute the p-value for the test of equality of survival curves as $p = 1 - F_{\chi_1^2}(0.505) = 0.477$, where $F_{\chi_1^2}(\cdot)$ denotes the cumulative distribution function of the χ_1^2 distribution. If our level of significance is set at $\alpha = 0.05$, then our conclusion could be stated as, we have insufficient evidence to conclude that the distribution of event times differs between arms A and B.

7.4.3 Proportional Hazards Model

The log-rank test is quite useful, but it has a few important limitations. First, flexible adjustment for a broad spectrum of covariates is not immediate. Second, the log-rank test does not furnish an easily interpretable summary of the size of difference between the survival curves. Third, extension of the log-rank test to quantitative covariates, such as age, is also unclear. The Cox proportional hazards model addresses each of these limitations of the log-rank test by extending the essential idea of conditioning on numbers at risk and total events for each time.

The proportional hazards model makes the assumption that the hazard for a subject with covariate vector x, $\lambda(t|x)$, is proportional to a baseline hazard $\lambda_0(t)$ times a positive constant $c(x)$. The constant is typically expressed as a log-linear function of the predictors, giving

$$\lambda(t \setminus x) = \lambda_0(t) e^{x_1\beta_1 + \ldots + x_p\beta_p} = \lambda_0(t) e^{x'\beta},$$

where the notation $x'\beta = x_1\beta_1 + \cdots + x_p\beta_p$ denotes the inner product of vectors (of the same length) x and β. Cox proportional hazards model fitting is based on a *partial likelihood*, which has conditioned out dependence on the baseline hazard $\lambda_0(t)$. The remaining parameters $\beta_1, \ldots, \beta_p$ each provide the change in log *hazard ratio* (HR) per 1 unit change in each predictor, while holding the others constant.

We consider a slightly simplified setting where *exactly one event* occurs at each unique *event* time, $t_1 < t_2 < \cdots < t_k$. It turns out that *tied* event times introduce a bit more complexity, which is not central to understanding how the proportional hazards model works. Approximations to the so-called *partial likelihood* in the relatively common case of tied event times are discussed briefly below. Further, consider a general multivariate regression setting where each observation is associated with a *vector of predictors* $x_i = (x_{i1}, \ldots, x_{ip})$. *Similar* to other regression models such as linear or logistic, categorical predictors with C levels need to be encoded as $C - 1$ linearly independent contrasts. Commonly, this is achieved by taking one of the category levels (say the "first") as the reference and representing category effects relative to the reference using *dummy variables*. For example, suppose we have a categorical variable with three levels A, B, and C. Then, we could set level A as the reference and represent this categorical variable using a 2-vector of predictors, $x_i = (0,0)$ if observation i has level A of the categorical variable, $x_i = (1,0)$ if level B, and $x_i = (0,1)$ if level C. Like other regression models, we can consider an interaction between two covariates, say x_1 and x_2, where the impact of each variable depends on the level of the other, by including their product $x_1 \times x_2$ in the model. *Unlike* other regression models such as linear or logistic, an intercept term (predictor variable always set to 1) is not included in a proportional hazards model. As we will see below, this "intercept" term is contained in the baseline hazard estimate. Statistical software for performing Cox proportional hazards modeling will typically handle the encoding of predictors automatically. Technical Material 666 provides a provides a brief construction of the Cox proportional hazards regression model.

Technical Material 666: Brief construction of Cox proportional hazards model. Recall the definition (Technical Material 333) of the *hazard function*, or force of mortality, as the event rate in the next small time segment conditional on a subject having survived up to a particular time of interest t. As noted above, the proportional hazards model makes the assumption

$$\lambda(t|x) = \lambda_0(t)e^{x_1\beta_1+\cdots+x_p\beta_p} = \lambda_0(t)e^{x'\beta}.$$

Consider two predictor vectors x and y that differ only in the jth component, with x_j 1 unit larger than y_j. Then the ratio of hazard functions is

$$\frac{\lambda(t/x)}{\lambda(t/y)} = e^{\beta_j} \text{ or } \log\frac{\lambda(t/x)}{\lambda(t/y)} = \beta_j.$$

Similar to the KM model, focus on a single event time t_k. Let $i(k)$ denote the index of the observation that has an event at time t_k and let R_k denote the set of indices of observations that are at risk just before the event at t_k occurs. Then, conditional on the risk set R_k, and conditional on a single event occurring at time t_k, the probability that the event occurs for observation $i(k)$ is the partial likelihood component k

$$L_k = \frac{\lambda\left(t_k|x_{i(k)}\right)}{\sum_{i\in\mathcal{R}_k}\lambda(t_k|x_i)} = \frac{\lambda_0(t_k)e^{x'_{i(k)}\beta}}{\sum_{i\in\mathcal{R}_k}\lambda_0(t_k)e^{x'_i\beta}} = \frac{e^{x'_{i(k)}\beta}}{\sum_{i\in\mathcal{R}_k}e^{x'_i\beta}},$$

which does not depend on the baseline hazard $\lambda_0(\cdot)$. Notice that if multiple events occur at time t_k, say m of them, then the numerator would be a product of m terms and the denominator would be a sum, over all possible ways of choosing m objects from the risk set R_k, of the products of m terms. This denominator can be very (and unnecessarily, given high-quality, computationally efficient approximations) computationally demanding, since $\begin{pmatrix} n_k \\ m \end{pmatrix}$ is very large for even moderate n_k. Typically, if tied event times are present, accurate and computationally efficient Breslow or Efron approximations to the likelihood are used [7]. Note that the number at risk n_k equals the size of the risk set R$_k$. The partial likelihood is the product of the above components over event times t_k,

$$L = L(\beta) = \prod_k \frac{e^{x'_{i(k)}\beta}}{\sum_{i\in\mathcal{R}_k} e^{x'_i\beta}}.$$

It can be shown that this partial likelihood enjoys many of the large sample properties of ordinary likelihood; in particular, large sample inference.

7.4.4 Testing and Confidence Intervals

In general, the maximum likelihood estimate of the vector of coefficients β cannot be expressed using a simple formula and must be found using iterative optimization techniques [8]. For large samples, the vector of coefficient estimates $\hat{\beta}$ is approximately normally distributed, centered around the true coefficients β and with variance-covariance matrix approximately equaling the (matrix) inverse of the *observed* Fisher's information matrix, $\mathrm{Var}\left(\hat{\beta}\right) \approx \left(-\mathbb{E}\left(\frac{\mathrm{d}^2}{\mathrm{d}\hat{\beta}^2}\log L\left(\hat{\beta}\right)\right)\right)^{-1}$.

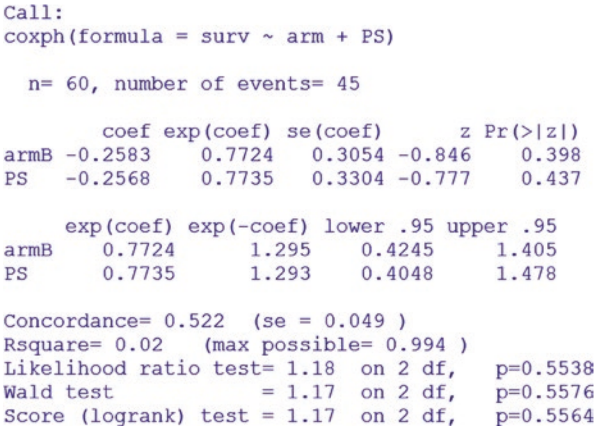
Notably, this variance-covariance matrix of the parameter vector can be extracted from most statistical software. This fact can be used to construct so-called *Wald* tests and confidence intervals for coefficients (log HRs) of interest. Suppose β_j is the parameter of interest. Let $\mathrm{SE}\left(\hat{\beta}_j\right) = \sqrt{\mathrm{Var}\left(\hat{\beta}\right)_{jj}}$ denote the standard error (estimated standard deviation) of the *j*th coefficient estimate $\hat{\beta}_j$, where $\mathrm{Var}\left(\hat{\beta}\right)_{jj}$ denotes the entry in the *j*th row and *j*th column of the variance-covariance matrix $\mathrm{Var}\left(\hat{\beta}\right)$. Then, a test of $\mathrm{H}_0 : \beta_j = \beta_j^0$ can be conducted by computing the test statistic $Z = \left(\hat{\beta}_j - \beta_j^0\right) / \mathrm{SE}\left(\hat{\beta}_j\right)$. Under the null hypothesis that $\mathrm{H}_0 : \beta_j = \beta_j^0$, Z is approximately distributed standard normal $\mathcal{N}(0,1)$ and an approximate (two-sided) p-value can be computed as $p = 2(1 - \Phi(|Z|))$, where Φ denotes the standard normal cumulative distribution function. A level $(1 - \alpha) \times 100\%$ confidence interval for β_j is given by $\hat{\beta}_j \pm z_{1-\alpha/2} \times \mathrm{SE}\left(\hat{\beta}_j\right)$.

Alternatively, likelihood ratio or score tests can be used to conduct hypothesis tests on the parameters β (or linear combinations of the βs). Notably, when there is only one binary predictor and a corresponding single β, the score test is exactly the log-rank test described above. For large samples, the likelihood ratio, score, and Wald tests are approximately equivalent.

7.4.5 Example

Consider again the example above where patients are randomly assigned to either arm A or arm B. In addition to the arms to which the patients are randomly assigned, each patient has a baseline performance status (PS), which measures the degree to which their disease interferes with the activities of daily life. Here, PS 0 indicates no disease interference, while PS 1 indicates that their disease interferes with strenuous activities. We could be interested in assessing the role of both arm and PS in the survival time. To set up the Cox proportional hazards regression, we need to encode both variables numerically and consider the interpretation of the encoding. Here, we consider arm A and PS 0 each as the *reference* category and encode both as 0, while arm B and PS 1 are encoded as 1. Once again, encoding of categorical predictors is generally done automatically by statistical software. With this encoding, β_1 is the log HR for arm B as compared with arm A (while holding the PS constant) and β_2 is the log HR for PS 1 as compared with PS 0 (while holding the arm constant).

Here, we consider the full sample of 30 observations each in arms A and B. Examine the R [2] output in Fig. 7.4, which is representative of statistical software output for Cox proportional hazards modeling. We see that, while there were 60 total observations, only 45 PFS events were actually observed. The remaining 15 PFS times were right-censored. The coefficient estimates (log HRs) $\hat{\beta}_1 = -0.258$ and $\hat{\beta}_2 = -0.257$ are provided along with their respective standard errors $\mathrm{SE}\left(\hat{\beta}_1\right) = 0.305$ and $\mathrm{SE}\left(\hat{\beta}_2\right) = 0.330$. These can be used to perform Wald tests of



```
Call:
coxph(formula = surv ~ arm + PS)

  n= 60, number of events= 45

         coef exp(coef) se(coef)      z Pr(>|z|)
armB -0.2583    0.7724   0.3054 -0.846    0.398
PS   -0.2568    0.7735   0.3304 -0.777    0.437

      exp(coef) exp(-coef) lower .95 upper .95
armB     0.7724      1.295    0.4245     1.405
PS       0.7735      1.293    0.4048     1.478

Concordance= 0.522  (se = 0.049 )
Rsquare= 0.02   (max possible= 0.994 )
Likelihood ratio test= 1.18  on 2 df,   p=0.5538
Wald test            = 1.17  on 2 df,   p=0.5576
Score (logrank) test = 1.17  on 2 df,   p=0.5564
```

Fig. 7.4 Cox proportional hazards modeling output of R statistical software

the impact of each of arm and PS (given that the other is held constant) by computing $Z_1 = \left(\hat{\beta}_1 - 0\right) / \mathrm{SE}\left(\hat{\beta}_1\right) = -0.258 / 0.305 = -0.846$ and $Z_2 = \left(\hat{\beta}_2 - 0\right) / \mathrm{SE}\left(\hat{\beta}_2\right) = -0.257 / 0.330 = -0.777$, along with corresponding *p*-values $p_1 = 2(1 - \Phi(|Z_1|)) = 2(1 - \Phi(0.846)) = 0.398$ and $p_2 = 2(1 - \Phi(|Z_2|)) = 2(1 - \Phi(0.777)) = 0.437$, as reported in the output. These could be interpreted as showing that we do not have strong evidence to suggest that the arm impacts survival after accounting for PS and we do not have strong evidence to suggest that the PS impacts survival after accounting for the arm. Overall tests that all model coefficients equal zero $H_0 : \beta_1 = 0, \beta_2 = 0$ are given at the bottom of the output in Fig. 7.4. Notice that the likelihood ratio, Wald, and score tests are approximately equivalent and none provides evidence that either arm or PS impacts PFS time.

From another perspective, our best respective estimates of the HRs for arm B as compared with arm A, for a fixed PS, and for PS 1 as compared with PS 0, for a fixed arm, are $e^{\hat{\beta}} = \left(e^{-0.258}, e^{-0.257}\right) = (0.772, 0.774)$. We could interpret these HR estimates as follows. For a fixed PS, we estimate that the hazard for death or progression (PFS) is 22.8% (1 − 0.772 = 0.228) lower in arm B than in arm A, while for a fixed arm, we estimate that the hazard for death or progression (PFS) is 22.6% (1 − 0.774 = 0.226) lower for PS 1 than for PS 0. However, these estimates are quite uncertain, as reflected by the wide 95% confidence intervals. For example, we are 95% confident that the true HR for arm B as compared with that for arm A for a fixed PS is in the interval [0.425, 1.405]. As described above, this interval can be constructed as $\exp\left\{\hat{\beta}_j \pm z_{1-0.05/2} \times \mathrm{SE}\left(\hat{\beta}_j\right)\right\} = \exp\{-0.258 \pm 1.96 \times 0.305\} = [0.425, 1.405]$.

7.4.6 Proportional Hazards Assumption

The Cox proportional hazards model makes the assumption that the hazard function, or force of mortality, for patients with a particular covariate value equals the product of two terms, one which depends only on time and one which depends only on the covariate. In particular, if the dependence on the covariate is log-linear, then $\lambda(t/x) = \lambda_0(t) e^{x_1\beta_1 + \cdots + x_p\beta_p} = \lambda_0(t) e^{x'\beta}$. As with all *models*, this model for the covariate-dependent hazard function is *wrong*, but it can be *useful* in many situations [9].

First, the score (or equivalent log-rank) test that all coefficients equal zero $H_0 : \beta_1 = 0, \ldots, \beta_p = 0$ (or more generally some given pre-specified values) is always *valid* in the sense that the *false-positive rate* is controlled at the specified level of significance. In particular, when the null hypothesis is true, the probability that the test rejects the null (a false positive) is approximately the level of significance, commonly denoted as α. When the proportional hazards assumption is (approximately) correct, then the score test is (approximately) *optimal*, in the sense that its power, or probability of rejecting the null hypothesis, is highest among all tests controlling the false-positive rate at the same level, and the same holds true for the asymptotically equivalent likelihood ratio and Wald tests. On the other hand, when the proportional hazards assumption is not approximately correct, these tests will suffer from a loss of efficiency and, in turn, power, to some extent. In the case where proportional hazards is *not* true, the coefficient estimates represent a *weighted average* of log HRs *over time*, which in many cases is still a sensible summary of how the force of mortality differs for different patients. For example, suppose that the log HR for two groups differs from log(2), at the beginning of follow-up, to log(1.25), at the end of follow-up. Then the estimated log HR comparing the two groups *under the proportional hazard assumption* might be log(1.7). Note that the *weighting* in the average depends on the actual underlying hazard rate, with higher hazard rate time increments receiving more weight and *vice versa*.

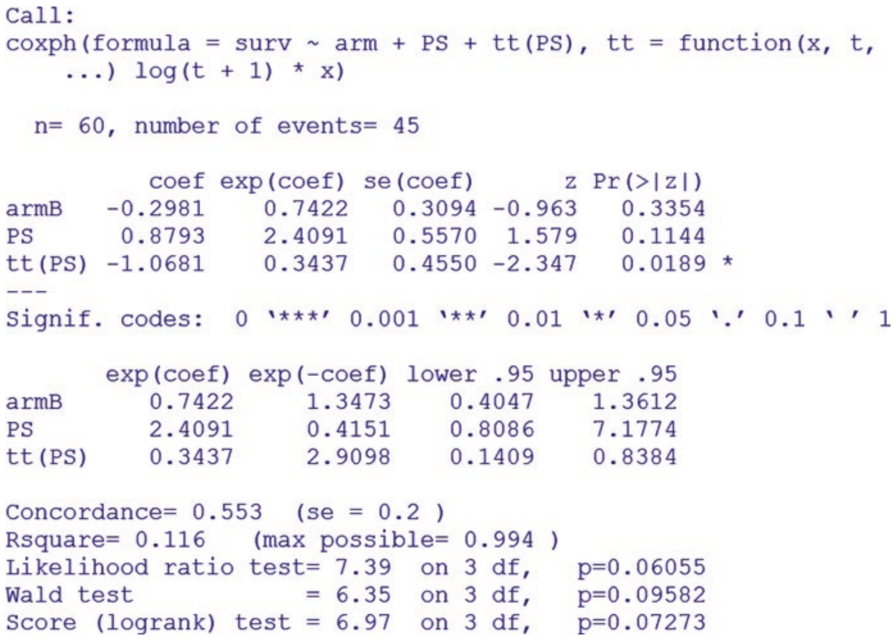
Diagnostics aimed at assessing the plausibility of the proportional hazards assumption generally fall into two categories, formal hypothesis tests of a covariate (or treatment) by time interaction or graphical assessments. To test whether a *particular* covariate β_j suffers from non-proportional hazards, a common approach is to introduce the auxiliary variable $z = \log(t + c) \times x_j$ and fit the model with $\lambda(t/x) = \lambda_0(t)e^{x_1\beta_1+\cdots+x_p\beta_p+z\gamma}$. Here, c is a fixed positive constant. Then, a test of $H_0 : \gamma = 0$ tests for an HR (due to changes in the jth covariate, given that all others are fixed) that is increasing or decreasing over time. If the estimate γ is negative this indicates that the HR is decreasing over time, while a positive value indicates that the HR is increasing over time. Note that this time-by-covariate interaction auxiliary variable is time-varying in the sense that its value for a particular patient is not fixed, but changes over time. Statistical software generally requires that time-varying covariates be handled in a different manner than covariates whose values are fixed across time.

Two common graphical approaches to assessing the proportional hazards assumptions check that so-called Schoenfeld residuals [10] do not have a systematic relationship with time and check that cumulative hazard estimates for groups of observations with similar covariate values are parallel. Schoenfeld residuals are defined as the difference between the observed and the expected covariate values for the patient with an *event* occurring at time t_k, $r_k = x_{i(k)} - \frac{\sum_{i\in\mathcal{R}_k} x_i e^{x_i'\beta}}{\sum_{i\in\mathcal{R}_k} e^{x_i'\beta}}$. Note that Schoenfeld residuals are not defined for censored observations. Further, at each event time, there is a vector of Schoenfeld residuals, one for each covariate in the model, say r_{kj} for the jth covariate. If the proportionality of hazards holds for the jth covariate, then there should not be a systematic relationship between the Schoenfeld residuals and the event times t_k when examining a plot of r_{kj} vs. t_k. A test of increasing or decreasing HR can even be conducted via simple linear regression of the Schoenfeld's residuals vs. time, for example. An alternative approach is to cluster the patients into groups that have similar covariate values, then for each group to compute and plot a cumulative hazard estimate over time (using Nelson-Aalen techniques as described above, for example). If proportional hazards are satisfied, then the cumulative hazard estimates should all be approximately parallel.

7.4.7 Example

Consider again our example. We could conduct a test for an HR *for PS* that is increasing or decreasing through time by including a (log) time-by-PS interaction in the model. Importantly, this time-varying covariate must be handled in a slightly different way than a fixed covariate in statistical software. Examine the R [2] statistical software output shown in Fig. 7.5. The *p*-value for the test that the HR for PS varies across time is 0.019, indicating strong evidence of non-proportionality of the HR for PS. The coefficient is negative (−1.068), indicating that the HR for PS 1 as compared with that for PS 0 (holding the arm fixed) decreases over time from 2.409 (= exp {0.879 − log (0 + 1) × 1.068}) at time 0 to 0.075 (= exp {0.879 − log (24.75 + 1) × 1.068}) at time 24.75.

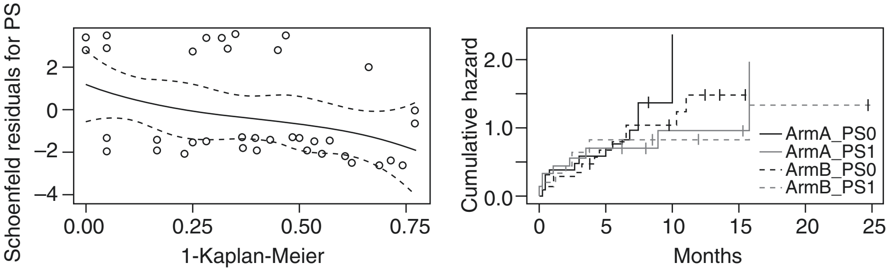
For a graphical perspective on the potential non-proportionality of the PS HR, see the plots in Table 7.4, where Schoenfeld residuals for PS are shown in the left panel and Nelson-Aalen cumulative hazard functions for each combination of arm and PS are shown in the right panel. Under the proportional hazards assumption, there should be no systematic relationship between the Schoenfeld residuals for PS and time. Notice that here time has been transformed to 1 minus the KM estimate $1-S(t)$, which tends to even out the distribution of times. Notice that the 95% confidence bounds containing the *smooth* (but not necessarily linear) relationship between the residuals and transformed time do *not* always contain the horizontal



```
Call:
coxph(formula = surv ~ arm + PS + tt(PS), tt = function(x, t,
    ...) log(t + 1) * x)

  n= 60, number of events= 45

           coef exp(coef) se(coef)      z Pr(>|z|)
armB    -0.2981    0.7422   0.3094 -0.963   0.3354
PS       0.8793    2.4091   0.5570  1.579   0.1144
tt(PS)  -1.0681    0.3437   0.4550 -2.347   0.0189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
armB      0.7422     1.3473    0.4047    1.3612
PS        2.4091     0.4151    0.8086    7.1774
tt(PS)    0.3437     2.9098    0.1409    0.8384

Concordance= 0.553  (se = 0.2 )
Rsquare= 0.116   (max possible= 0.994 )
Likelihood ratio test= 7.39  on 3 df,   p=0.06055
Wald test            = 6.35  on 3 df,   p=0.09582
Score (logrank) test = 6.97  on 3 df,   p=0.07273
```

Fig. 7.5 R statistical software output for Cox proportional hazards model with time-varying performance status (PS) effect

Table 7.4 Schoenfeld residuals for performance status (PS) are shown in the left panel. Cumulative hazard estimates by covariate combination are shown in the right panel



line $y = 0$, and we, in turn, might conclude that there is evidence of non-proportional hazards for PS. The Nelson-Aalen cumulative hazard estimates are roughly parallel up to around time 7 months, after which the hazard appears to accumulate more slowly in the PS 1 groups (gray curves) as compared with the PS 0 groups (black curves).

7.4.8 Predicting Survival Probabilities with the Cox Model

The Cox proportional hazards model was explicitly constructed to remove dependence on the baseline hazard function. If our only interest is in assessing the influence of variables in terms of shifting the hazard up or down, then this is an advantage, since it does not require estimating the potentially very complex baseline hazard function, which would be associated with an increase in our level of uncertainty (variance) about the influence of these variables of interest. In many situations, however, there is a direct interest in making survival probability predictions for particular types of patients, and these survival predictions, in turn, depend on the baseline hazard. For example, referring to our continuing example, we may want to predict the 2-year survival rate for PS 1 patients receiving the treatment in arm A. A straightforward approach in this situation would be to construct a KM survival curve estimate just using the PS 1 patients in arm A, but this approach does not simply extend to covariates such as age. We may want to generate a survival prediction for a 45-year-old patient, but our data may contain no or relatively few 45-year-olds. Instead, we would like to use a *model* to predict the survival.

Most statistical software can generate these types of predictions, which are (typically) conditional on the parameter estimates of the Cox proportional hazards model and involve the additional step of estimating event probabilities at each unique event time. There are two very important complexities to consider when generating these types of model-based survival probability predictions. First, the model needs to be reasonably reflective of the data for the predictions to be meaningful. Some common reasons why the model might not fit the data well include non-proportional hazards, a non-log-linear relationship between the covariates and hazard function, or leaving an important covariate out of the model, perhaps because the important covariate was not measured or is not present in the data, for whatever reason. Second, model-based survival predictions typically require the complete specification of all covariates included in the model, as well as a specific time of interest. For example, if age, PS, and treatment are included in the model, then predictions can only be generated for a specified age, PS, and treatment. Most statistical software will also require that a specific time of interest is specified. Not all statistical software handles survival predictions the same way by default. A common default behavior is to make a survival probability prediction for each observation in the data set at its observed time (whether a censoring or event is observed). If one would like to predict the survival probability at a time of interest *averaged over the distribution of some covariate(s)*, gender and age-averaged, for example, then predictions would need to be generated for the range of covariate combinations and then combined using a weighted average with weights corresponding to the relative frequencies of the combinations. In particular, it does not make sense to consider an average covariate value for most categorical variables. For example, it does not make sense to make a prediction for the average gender in a sample.

7.4.9 Variable Selection, Model Building, and Stratification

In many situations where we have several potentially important predictor variables, we would like to construct a model that appropriately accounts for the influence of all these variables. This could be useful for prediction or for adjusted effect estimates where we estimate the impact of changing one variable while the others are held constant. A common approach to building a multivariate model is *forward step-wise variable selection*. This approach works by starting from a model with no covariates, then adjusting the current model step by step by either adding or removing variables (or taking no action) according to some criterion. When the criterion for taking no action is best in the current model, the procedure stops. An alternative approach is *backward step-wise variable selection*, where the starting model has all candidate variables, although this approach is not well-suited to situations with a large number of potential covariates. Note that these step-wise variable selection procedures can allow one to entertain non-log-linear relationships between covariates and hazard by including non-linear transformations of variables, say age^2 or log(age), additionally as candidate variables.

Reasonably well-established, and relatively high-quality, criteria for variable selection include hypothesis testing, information criteria such as the Akaike [11] and Bayesian [12] information criteria (AIC and BIC), and cross-validation. Hypothesis testing can be used by setting entry and exit p-value thresholds, such as <0.05 and ≥ 0.05, for including and removing variables. AIC and BIC work by penalizing the (log) likelihood according to the number of parameters in the model. Note that if the current model is increased in complexity, say by including a new variable, the likelihood can only increase. The more complex model is better able to accommodate the variability in the data. On the other hand, some aspect of the variability in the data is not actually related to the covariates, but instead is just random *noise*. We want to fit the data well, but not *over-fit* the data (fit model to random noise in addition to the signal). BIC and AIC, respectively, penalize additional parameters more and less strongly, making BIC tend to give a simpler model and AIC a more complex model. Both AIC and BIC are *the smaller the better* criteria. Cross-validation works by randomly dividing the data in *training* and *testing* sets several times. For each division into training/testing, the model of interest is fitted using the training data; then its performance is measured on the testing data, according to some measure, partial likelihood, for example. Typical divisions of the data into training/testing include leave-one-out (one testing case withheld) and fivefold and tenfold (1/5 and 1/10 of data withheld for testing).

It is important to make a few notes on the prediction from and interpretation of multivariate models. The above-described techniques for building a multivariate model are purely data-based (as opposed to relying on expert knowledge). If the model fits (not over-fits) the data well, then predictions might be expected to perform reasonably well for new data that *is similar to the data used to build the model*. In general, extrapolating predictions beyond the range of the observed covariates is not advisable, but typically the situation is even more complex than simple extrapolation due to *collinearity*, or strongly related prediction variables. If two variables are very strongly related, say they are both large or both small together, then the data does not inform predictions for when one is large and the other small and *vice versa*. Making a prediction for this type of *unobserved combination* is, in a sense, extrapolation. Interpretation is also very challenging in the presence of collinearity. For some closely related variables, say two different pain scores, it does not even make sense to talk about changing one variable while the other is held constant. Even if it makes sense to talk about changing only one variable while the other is constant, the data may not offer much information, and, in turn, the estimates of the effect of changing only one variable are not trustworthy. Still another complexity is in assessing which of a group of closely related variables is most important. If the variables are very closely related, a data-driven approach, such as those described above, cannot reliability identify which variable is most closely associated with event times.

In some situations, there is a very important categorical variable or grouping which, while not of direct interest, needs to be adjusted for in a manner that does not require proportionality of hazards. For example, in a multicenter trial, we might expect the hazard functions to differ by site in a manner that is potentially more complex than proportional. We are not typically interested in the center effects *per se*, but would like to adjust for differences between sites. In a situation like this, one could fit a Cox proportional hazards model *stratified* by site. A Cox proportional hazards model with a stratification factor conditions out the effect of the factor by allowing a distinct baseline hazard function for each stratum. The impact of the remaining factors is then modeled as constant across all strata. The HR per unit change in the non-stratification variables is the same for each stratum.

7.4.10 Power, Sample Size, and Follow-Up

As we saw in Chap. 5, power and sample size are important considerations for study planning and the interpretation of non-significant hypothesis tests. For survival analysis, the relevant aspect of *sample size* is the *number of events*. A very useful formula relating the total number of events D, level of significance α, power $1-\beta$, respective proportions of patients in groups A and B, p_A and p_B $(=1-p_\mathrm{A})$, and the HR for the groups of interest is Schoenfeld's formula [13]

$$D=\frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{p_\mathrm{A}p_\mathrm{B}\left(\log \mathrm{HR}\right)^2}.$$

Written in terms of power, this equation becomes

$$1-\beta=\Phi\left(\sqrt{Dp_Ap_B}\left|\log \mathrm{HR}\right|-z_{1-\alpha/2}\right),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (recall that $z_{1-\beta}=\Phi^{-1}(1-\beta)$, the inverse of the standard normal cumulative distribution evaluated at $1-\beta$).

For example, if we want 80% power $(1-\beta=0.80)$ with a level of significance of 0.05 $(\alpha=0.05)$ in a randomized trial with equal allocation to each arm $(p_\mathrm{A}=p_\mathrm{B}=0.5)$, under the assumption that the HR is 0.75 (HR = 0.75), then approximately

$$D=\frac{\left(z_{1-0.05/2}+z_{0.80}\right)^2}{0.5\times0.5\times\left(\log 0.75\right)^2}=\frac{\left(1.96+0.84\right)^2}{0.5\times0.5\times\left(\log 0.75\right)^2}\approx 380$$

events are required. From the other perspective, suppose 500 events are observed $(D=500)$ with one-third of patients in group A $(p_\mathrm{A}=1/3)$ and two-thirds in group B $(p_\mathrm{B}=2/3)$ with a level of significance of 0.05 $(\alpha=0.05)$. If the true HR is 1.35, then the test of group A vs. group B has power

$$1-\beta=\Phi\left(\sqrt{500\times(1/3)\times(2/3)}\left|\log 1.35\right|-1.96\right)\approx 0.89.$$

An implication of this dependence of power on the number of events is the importance of *sufficient follow-up*. The patients must be followed for long enough that the required number of events occurs.

7.5 Additional Topics

7.5.1 Parametric Regression

As we discussed before, outcomes of interest in many cancer studies are times-to-event, but often we have additional data (e.g., age or ethnic group) and are interested in investigating the effects of these covariates on the survival distribution. One possible regression modeling framework that we discussed before is given by the Cox proportional hazards model. An alternative approach to incorporate covariates is offered by parametric models. As we have seen before, parametric methods require assumptions about the underlying form of the survival distribution, but they can achieve a more precise estimation of the parameters if they fit the data well.

Parametric models that examine how predictors affect the hazard function include parametric proportional hazards models and additive hazards models. Parametric proportional hazards models describe the dependency between the covariates and failure time in terms of the HR and are similar to the Cox proportional hazards model, with the only difference being that the baseline failure rate is assumed to be parametric. The choice of the parametric distribution is based on model fit. To accommodate for the fact that event times are positive and are generally skewed, common choices for a parametric distribution include Weibull, exponential, log-normal, and log-logistic.

While parametric proportional hazards models include covariate effects in a log-linear manner, additive hazards regression models assume additive covariate effects, leading to an easier interpretation of the effects of the covariates. Additive hazards regression models investigate the relationship between the covariates and failure time in terms of the change in hazard function owing to the exposure of interest, which can be thought of as an attributable risk due to exposure.

While the two parametric models discussed above model the effect of a set of predictors on the hazard function, accelerated failure time models (AFTs) examine how covariates affect the survival function directly. Let T denote a random time-to-event. The AFT model regresses the logarithm of the survival time on the predictors, and the general framework of the model assumes the following form:

$$\log T = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon,$$

where $x_1, \ldots, x_p$ are the predictors, $\beta_1, \ldots, \beta_p$ are the parameters, and ϵ is the error. While the AFT model accommodates censored observations, in the absence of censoring, the above setup translates into an ordinary least squares regression model. Similar to before, the distribution of the error has to be pre-specified by the investigator. The parameters $\beta_1, \ldots, \beta_p$ each provide the change in the log *time ratio* (TR) per 1 unit change in each predictor while holding the others constant. TR can be thought of as the acceleration factor. TR > 1 (TR < 1) implies that it takes more (less) time for an event to occur, which means that an event is less (more) likely to occur.

Although the AFT model is not as widely used as the other models, it is a good alternative approach because it is less affected by the choice of the error distribution, and the results are easily interpreted.

7.5.2 Landmark Analysis

We saw above that the outcome of interest in many cancer studies is time-to-event. Another frequently used endpoint is the objective tumor response. There are four categories of tumor response: (1) complete response, (2) partial response, (3) stable disease, and (4) progressive disease. Responders are defined as those patients who have either a complete or a partial response. The last two categories define non-responders. A common practice in survival analysis is to compare the survival between responders and non-responders.

As an example, consider a case where responders survive significantly longer than non-responders. Comparing the survival curves for the two groups, one might conclude that the effect of response is to extend life. This approach is problematic because it introduces bias. A patient belongs to the responder group only if s/he survived until the time of response evaluation and was evaluated as a responder. Responders are guaranteed a survival time that is at least as long as the time to the first response evaluation, while patients who die before the first evaluation are automatically labeled non-responders. This approach is biased in favor of responders, while producing an incorrect unfavorable survival curve for non-responders. Moreover, the guarantee time for responders provides a better chance for the therapy to produce a response. This bias, caused by the guarantee time for responders, results in invalid statistical comparisons of the survival distributions of responders and non-responders. A technique called landmark analysis addresses this issue and corrects the bias.

A landmark analysis places a fixed time after the initiation of treatment as a landmark for conducting the analysis of survival by response. Patients who are not alive at the landmark time are excluded from the analysis. Those who are alive at the landmark time are separated into two response categories according to whether or not they have responded to the treatment up to that time. Survival is credited from the time of the landmark, and patients are analyzed according to their response status at the landmark time, regardless of any changes in the response status after that. The conditional nature of the landmark analysis removes the previously discussed bias by assigning each patient to a response group at the landmark time and estimating the survival probabilities as functions of response status at that baseline.

Although this approach results in a correct statistical test, it has several limitations. The main disadvantage of the method is its sensitivity to the choice of landmark time; results might be different depending on the selected landmark time. Furthermore, in order to avoid additional biases, the landmark time should be selected prior to the data analysis and based on some clinically significant natural time. Another problem is the omission of events occurring earlier than the landmark time, and the omission of shifts in response status after the landmark time. Finally, the landmark method does not address the bias in the formation of the groups based on the outcome. This lack of randomization is problematic, since in many cases responders are the patients with better prognostic characteristics.

7.5.3 Recurrent Events, Competing Risks, Interval Censoring, Informative Censoring, and Dependent Observations

There are a handful of important topics (mentioned in the title of this subsection) that are beyond the scope of this chapter to discuss in detail, but that warrant a few comments and references. In general, each of these situations requires specialized techniques and interpretation.

Recurrent events occur when a particular type of event can happen to a patient repeatedly; for example, cancer recurrences. Common models for recurrent events include the Andersen-Gill, Prentice-Williams-Peterson, marginal rates, frailty, and multi-state models. A useful introduction to recurrent events modeling is given in [14]. Competing risks occur when a patient is subject to several risk processes simultaneously. For example, cancers, particularly those that occur in elderly patients and are associated with indolent behavior, e.g., hormone-sensitive prostate cancer, are commonly subject to recurrence and mortality risk processes at the same time. Common models for competing risks include the Cox proportional hazards, Fine and Gray, multi-state, and inverse probability of censoring weighted models. A useful introduction to competing risks modeling is given in [15].

Interval censoring occurs when the event of interest is not known exactly, but is known to have occurred within a time interval. In fact, right-censoring is a special case of interval censoring, where the event of interest is known to occur between a time of interest and infinity (the interval is open on the upper end). Common techniques relating to interval-censored data include the self-consistency/Turnbull method, extensions of the traditional Cox proportional hazards model for right-censored data, and parametric models. A useful introduction to modeling interval-censored data is given in [16]. Most of the techniques for censored data require that event times are independent of censoring time. If this assumption is not satisfied, say for TTP subject to censoring by death, then most analyses are invalid, including KM and Cox proportional hazards analyses. Techniques for handling informative censoring include multiple imputation [17] and inverse probability of censoring weighting [18].

Dependent observations occur when the event times for some groups of patients tend to be more similar than the event times for patients in different groups. For example, the event times for patients treated by the same doctor may tend to be more similar than the event times for patients treated by different doctors. A common technique for handling dependent observations is to use a so-called robust (sandwich) estimator of variance, which adjusts the estimates of parameter uncertainty to account for a grouping structure [19].

Conclusion

Survival is the gold standard primary endpoint in oncology clinical research because it reflects an undeniable benefit to cancer patients. Understanding the basic concepts involved in the analyses of survival is quite important for interpreting the literature and/or designing trials with time-to-event endpoints. This chapter has highlighted some of the challenges faced when one is summarizing and analyzing time-to-event data, and has also discussed some of the common approaches to handling these challenges appropriately. Several of the more complex challenges, such as recurrent events, competing risks, interval censoring, informative censoring, and dependent observations, have been discussed only very briefly, with references provided for further study.

References

1. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. New York: Springer Science & Business Media; 2006.
2. R Core Team. A language and environment for statistical computing. Vienna: R Core Team; 2014. http://www.R-project.org/
3. Moore D. Applied survival analysis using R. New York: Springer; 2016.
4. Van Cutsem E, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. N Engl J Med. 2009;360(14):1408–17.
5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.
6. Greenwood M Jr. The natural duration of cancer. Reports of Public Health and Related Subjects. London: HMSO; 1926. p. 33.
7. Efron B. The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc. 1977;72(359):557–65.
8. Jorge N, Stephen JW. Numerical optimization. New York: Springer; 1999.
9. Box GEP. Robustness in the strategy of scientific model building. Robustness Stat. 1979;1:201–36.
10. Hosmer DW, Lemeshow S. Regression modeling of time to event data, Applied survival analysis. New York: Wiley; 1999.
11. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–23.
12. Schwarz GE. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
13. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983;39:499–503.
14. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33.
15. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389–430.
16. Lindsey JC, Ryan LM. Methods for interval-censored data. Stat Med. 1998;17(2):219–38.
17. Sterne JAC, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
18. Van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer Science & Business Media; 2003.
19. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. J Am Stat Assoc. 1989;84(408):1074–8.

The Role of Cross-Sectional and Cohort Studies in Oncology

8

André Lopes Carvalho, Fabiana de Lima Vazquez,
and Cleyton Zanardo de Oliveira

8.1 Introduction

In this chapter, we discuss the role of cross-sectional and cohort studies in oncology. We will particularly emphasize the definitions, uses, calculations, limitations, and the advantages and disadvantages of these study designs. Examples will be given to better illustrate the definitions and context.

In other words, we will describe what is important to consider when reading scientific information obtained from these types of studies, including issues with sample size, biases, internal and external validity, and completeness of follow-up information. We hope this information will help the reader to understand better and to critically interpret the results described in the literature.

8.2 Cross-Sectional Studies

Cross-sectional studies are classified as observational studies of an analytical nature. In analytical studies, we compare groups to analyze the effect of a particular exposure in a group of people compared with a group of unexposed individuals, and we investigate certain outcomes.

In cross-sectional studies, information on exposure is collected together with information on outcome. Thus, it is very important to emphasize that, in this type of study, the disease occurrence, exposure, and event are measured simultaneously at a single point in time (or in a short period). There is no follow-up period with the

A. L. Carvalho, M.D., Ph.D., M.P.H. (✉) • F. de Lima Vazquez, D.D.S., Ph.D.
Teaching and Research Institute, Barretos Cancer Hospital, Barretos, SP, Brazil

C. Z. de Oliveira, M.Sc.
Teaching and Research Institute, Barretos Cancer Hospital, Barretos, SP, Brazil

Education and Research, BP - A Beneficência Portuguesa de São Paulo, São Paulo, SP, Brazil

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_8

study participants; thus, what we obtain is a “picture” of the reality to be studied. This helps us to understand what happens with a certain population at that point in time and to infer possible associations from the results. As there is no temporal follow-up in such studies, we cannot establish the causality or association between exposure and the subsequent development of the disease under investigation, because we cannot determine what has occurred first (exposure-disease/cause-effect). The main measure of a cross-sectional study is prevalence.

Cross-sectional studies should aim to estimate the prevalence of the outcome of interest in a given population, and this is the type of study indicated for public health planning. These studies are impractical for investigating rare diseases, which have a low prevalence, as they would then require very large samples.

However, there are advantages of cross-sectional studies; they are fast paced, and rapidly provide information, results, and conclusions from data analysis. Also, this type of study has a low cost and is simple from the analytical/statistical point of view. The lack of a follow-up period eliminates problems associated with follow-up losses, which are common in cohort studies. In short, cross-sectional studies are faster, cheaper, and logistically simpler than cohort or case-control studies; cross-sectional studies are appropriate for measuring prevalence, and allow hypothesis generation regarding causality.

8.2.1 Applicability

Cross-sectional studies are suitable for the characterization of a population in relation to certain variables and their distribution. They have a descriptive purpose, and are often carried out in the form of a survey, from which we can obtain useful data to estimate or assess health needs in a given population.

There is usually no defined hypothesis; the purpose of these studies is to describe a population or subgroup within the population in relation to a result and a set of risk factors, and to find the prevalence of the outcome of interest (e.g., disease) within the population, or population subgroups, at any given time.

Example: A cross-sectional study was conducted to find the prevalence of human papillomavirus (HPV) genotypes in cervical cancer in samples from 38 countries. This study reported which types of HPV should be prioritized when evaluating the cross-protection effects of current vaccines and formulated recommendations for the use of second-generation HPV polyvalent vaccines. The study demonstrated that the prevalence of HPV genotypes varied with geographic location. Three genotypes (HPV types 16, 18, and 45) were the most prevalent in any studied population and it was concluded that these genotypes should be the focus for screening tests and clinical protocols [1].

Cross-sectional studies also allow the use of repeated cross-sectional surveys, with random sampling and standardized settings, which provide useful trend indicators.

Example: A study with data from 2009 to 2013 referred to cytological examinations that were collected from a cervical cancer screening system in Brazil (SISCOLO). Cytopathological examinations (n = 62,397,698) from women aged between 25 and 64 years were collected at SISCOLO, and studied according to the collection site and age of the women who voluntarily participated in the opportunistic cervical cancer prevention program of the Brazilian Government. The annual

Table 8.1 Values of the cervical cancer screening program quality indicators in Brazil 2006–2013

	Year							
Indicator	2006	2007	2008	2009	2010	2011	2012	2013
Productivity rate (%)	15.98	16.72	16.72	17.32	16.61	15.97	15.51	–
% Exams performed	74.69	73.65	76.62	77.00	77.48	77.96	78.36	78.69
% Unsatisfactory	1.07	1.10	1.02	1.10	0.93	0.91	0.95	0.99
% Rejected	0.10	0.11	0.12	0.11	0.12	0.17	0.24	0.29
% TZ (≥50 years)	54.11	52.85	49.18	47.13	46.00	45.12	45.71	46.03
% TZ (<50 years)	66.83	67.74	66.92	65.37	65.12	64.07	64.91	64.94
% Positivity Index	2.64	2.57	2.50	2.48	2.62	2.64	2.59	2.72
%ASC-US	1.15	1.08	1.11	1.15	1.25	1.24	1.25	1.27
% ASC-H	0.17	0.18	0.18	0.18	0.20	0.21	0.22	0.24
% LSIL	0.82	0.76	0.69	0.64	0.66	0.67	0.63	0.54
% HSIL	0.33	0.31	0.30	0.29	0.30	0.30	0.30	0.27
% ASC	1.26	1.26	1.29	1.33	1.45	1.45	1.47	1.51
ASC/abnormal rate (%)	47.77	49.02	51.78	53.76	55.19	54.77	56.58	55.52
ASC/SIL ratio	1.09	1.17	1.31	1.42	1.51	1.49	1.57	1.87

doi:10.1371/journal.pone.0138945.t003

TZ transformation zone, *ASC-US* atypical squamous cells of undetermined significance, *ASC-H* atypical squamous cells cannot exclude high-grade squamous intraepithelial lesion, *ASC* atypical squamous cells, *LSIL* low-grade squamous intraepithelial lesion, *HSIL* high-grade squamous intraepithelial lesion

percent change (APC) for each variable was evaluated, and change in trend was observed in four quality indicators, leading to the conclusion that the evaluation of the indicators from 2006 to 2013 suggested that actions should be taken to better control cervical cancer in Brazil. The data demonstrated a significant declining tendency of low-grade squamous intraepithelial lesions (LSILs) and high-grade squamous intraepithelial lesions (HSILs), and an increasing rate of invalid tests from 2009 to 2013. The number of positive cytological diagnoses was lower than expected, since developed countries with a low frequency of cervical cancer detect more lesions per year. The trend of the indicators during this period suggested that public health actions should be adopted to improve the effectiveness of cervical cancer control in Brazil [2] (Table 8.1).

Note that although the study period was seven years, this was not a cohort study but a cross-sectional one, as the information on each variable was collected at a single point in time and the women evaluated were not the same for all data collected, but belonged to the same population.

8.2.2 What to Consider in a Cross-Sectional Study

The design of a cross-sectional study must include certain parameters, such as the precise identification of the question to be answered, the studied population, whether the study uses census or sampling, the presence or absence of outcome and exposure for all individuals, and methods used to measure the variables of interest.

8.2.2.1 Population and Sampling

Sampling

Sample selection and response rate determine to what extent the results can be representative, or even generalized, to the entire population. Thus, sampling should consider that all individuals in the study population have a similar chance of being included in the study.

Selecting the population of the study depends on the objective, or what the study aims to investigate, such as, for example, the population of a city/by age group/women, or people followed at a cancer hospital/by workers/work-related exposure. Moreover, sample size should be sufficient to estimate the prevalence of the condition of interest with adequate accuracy. For details about how to calculate sample size for observational studies see Chap. 5.

In a cross-sectional study, much information on potential risk factors can be collected, allowing hypothesis generation for various outcomes and exposures (Fig. 8.1).

The sample should represent the population, considering that:

- All individuals in the target population should have the same chance of entering the study as if they were randomly selected to participate.
- The sample size must be representative.
- If convenience samples are used, the methodology should be well-described and justified.

Example: In a cross-sectional study conducted between 2008 and 2011, prostate glands were prospectively collected from 320 men within 24 h of death to compare the prevalence of prostate cancer (PCa) in a specific population of Caucasian and Asian men who died of causes unrelated to prostate cancer. PCa prevalences of 37.3% among Caucasians and of 35% among Asians were observed [3].

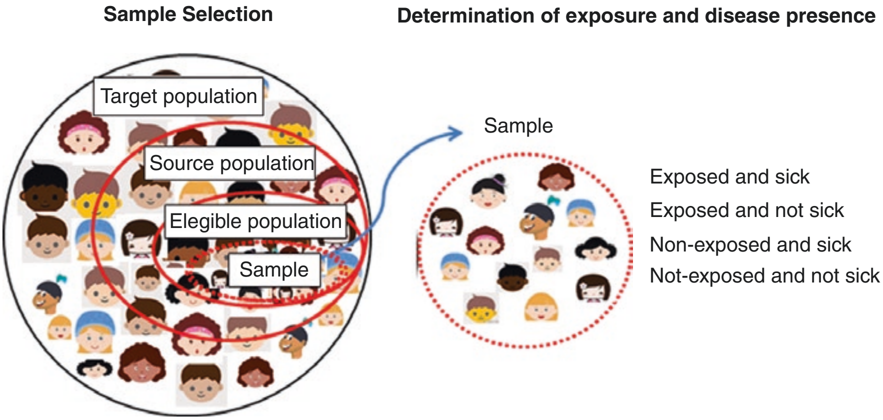


Fig. 8.1 Sample selection, determination of exposure, and disease presence

Table 8.2 Examples of target population, source population, eligible population, and sample

Target population	For whom?	I want to extrapolate the results	Asian and Caucasian men
Source population	For whom?	The study can lead to inferences	Asian and Caucasian men who died of causes other than prostate cancer (PCa)
Eligible population	Who is?	Eligible for study	Men who died from causes other than PCa in Moscow, Russia (Caucasian), and Tokyo, Japan (Asian) between 2008 and 2011 and aged 20 to more than 80 years
Sample	Who is it?	Enrolled in the study	Men (*n*=320, who died of causes other than PCa

Table 8.2 provides details of how to consider the sampling and the population that a cross-sectional study is trying to represent.

If the sample indeed represents the target population, which is not so easy, then an external validation is possible; that is, there is the possibility of generalizing the results obtained. Internal validation is given when the sampling represents the source population.

8.2.2.2 Frequency Measures in Observational Studies

Frequency measures indicate an occurrence and describe the distribution of events in absolute numbers, identifying groups at risk and suggesting hypotheses for a given problem, disease, or health condition. This measure is characteristic of cross-sectional studies, also known as prevalence studies. The prevalence of a particular type of cancer is defined by the proportion of diseased individuals in a population at a given time point.

Knowing this prevalence is of great importance for the implementation of prevention, treatment, and public health planning measures. The event or outcome must be clearly defined and characterized.

Exposure measures describe factors, variables, or measures that a person or a group of people has been exposed to and that may be relevant to the event or outcome, such as cigarette consumption or lifetime drinking.

Confounding measures are errors in the selection of individuals for the study or in the measurement of a variable. They distort measures of disease occurrence by over-estimating, under-estimating, or simply suggesting a false association between two variables. These types of errors are called bias, and can occur in the selection process (sampling) and in information collection (recall bias, interviewer bias, instrument bias, detection/diagnosis bias). Reverse causation is another cause of bias.

The most likely bias in a cross-sectional study is prevalence bias, which occurs when a factor related to disease duration is confused with the occurrence of the disease.

Example: A study conducted in the 1970s reported a high frequency of A2 human lymphocytic antigen (HLA-A2) among children with acute lymphocytic leukemia (ALL), concluding that children with HLA-A2 had an increased risk of developing this disease. Subsequent studies, however, demonstrated that HLA-A2 was not a risk factor for ALL, but rather, a factor associated with better prognosis. The longer survival of HLA-A2 children included in the cross-sectional study sample was associated with a greater probability of finding patients (survivors) with this type of HLA compared with other types [4].

Information bias may occur through the use of non-standard or non-validated methods, which provide inaccurate information. Examples involve the use of non-validated questionnaires or laboratory test analysis with inappropriate techniques. Information bias also includes recall bias, which refers to when a study participant has difficulty remembering past events, experiences, and exposures. The recall of such items may also differ between individuals who present the outcome and those who do not, and affected individuals may strive more to remember than non-affected individuals, imposing a bias on the information collected. Also, in a questionnaire, the interviewer should always be aware that respondents tend to provide socially accepted responses and not necessarily the truth, imposing a bias in the measures.

Reverse causality occurring in cross-sectional and case-control studies happens when the exposure changes as a consequence or result of the disease.

Example: The occurrence of tuberculosis and lung cancer in the same patient, concomitantly or not. It is possible that tuberculosis is more often diagnosed before lung cancer because of a reverse causality bias, i.e., hidden lung cancer may reduce immunity and lead to the reactivation of latent tuberculosis. Thus, tuberculosis may be clinically present before the clinical confirmation of lung cancer [5–8].

8.2.3 Data Analysis in Cross-Sectional Studies

Descriptive analysis: Only one variable of interest, and the average, median, standard deviation, etc., are described. This is a descriptive analysis of the data.

Analytic analysis: When two variables and the relationship between them are analyzed at the same time. Because cross-sectional studies are prevalence studies, we use the prevalence ratio to test statistical associations, i.e., the ratio between the number of diseased individuals and the population. A cross-sectional study also allows multivariate analyses.

Prevalence ratio: The prevalence ratio does not establish the risk of developing a disease, but rather, estimates how prevalent diseased individuals are among the exposed group compared with the prevalence in unexposed individuals (Table 8.3).

For example; in the cross-sectional study of HPV genotypes mentioned above [1], women who were HPV16/18-positive had a prevalence of cervical intraepithelial neoplasia (CIN)2+ that was 4.59 times higher than that in HPV16/18-negative women.

Table 8.3 Relationships between exposure and outcome in a cross-sectional study

	Outcome		
Exposure	**Yes**	**No**	
Yes	Individuals exposed with outcome	Individuals exposed without outcome	
No	Individuals not exposed with outcome	Individuals not exposed without outcome	
	Disease	**No Disease**	**Total**
Exposed	**a**	**b**	**a + b**
Not-exposed	**c**	**d**	**c + d**
Total	**a + c**	**b + d**	**N**
Example:			
Frequencies	**CIN2+**	**<CIN2**	
HPV16/18-positive	**153**	**134**	**287**
HPV16/18-negative	**111**	**842**	**953**
Total	**264**	**976**	**1240**

Prevalence (exposed): $a/(a + b) = 153/(153 + 134) = 0.533$
Prevalence (not-exposed): $c/(c + d) = 111/(111 + 842) = 0.116$
Prevalence ratio (PR) = prevalence (exposed)/prevalence (not exposed) = 4.59
HPV human papilloma virus, *CIN* cervical intraepithelial neoplasia

8.2.4 Advantages and Disadvantages of Cross-Sectional Studies

Cross-sectional studies, despite being easier to execute are modest in providing a high level of scientific evidence when compared to the other analytical studies. In Table 8.4 we can appreciate its advantages and disadvantages.

8.3 Cohort Studies

8.3.1 Definition

In epidemiology, a cohort is defined as a group of individuals who share a common characteristic or experience within a defined period (e.g., are exposed [or not] to a risk or prognostic factor, or undergo a certain medical procedure). There is also the "birth cohort", defined as a group of people who were born in a particular period. Cohorts are bound together by similar circumstances and followed by a determined period of time.

There are some studies in oncology that describe the characteristics and the occurrence of an outcome in a particular cohort of individuals, without using measurements of association. However, here we will describe "cohort studies", which means an epidemiological study design to measure association in a cohort, explicitly comparing exposed and not-exposed subgroups within that cohort.

Cohort studies are observational, analytical, and longitudinal studies in which, from a potential baseline cause (exposure) one investigates its effect (outcome). Therefore, these studies represent a very good design for evaluating causality. They

Table 8.4 Cross-sectional studies: advantages and disadvantages

Advantages	Disadvantages
• Measure prevalence	• Do not measure incidence
• Low cost	• Not fit for investigations of low-prevalence outcomes
• Easy to conduct/simple logistics	• Do not determine cause and effect (causality)
• Rapid collection of data/data collected only once for each	• Temporality—do not determine duration of illness
• Real-time information	• No inference of statistical power
• No 'lost to follow-up'	• Susceptible to bias in selection and information, etc
• Monitor population health	• Do not assess exposure duration (risk or protection?)
• Aid in planning for public health	• Do not determine absolute risk
• Generate hypotheses	• Susceptible to bias owing to low response
• Provide inference for a defined population	

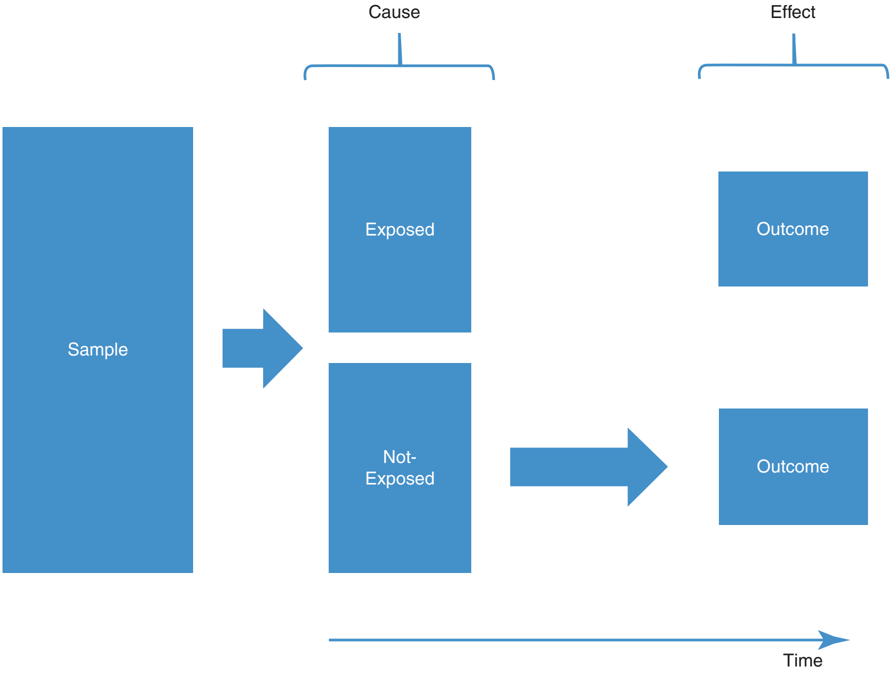


Fig. 8.2 Scheme depicting the design of a cohort study

can be prospective or retrospective and the study population is divided into subgroups according to exposure characteristics (exposed and not-exposed). The study then assesses exposed and not-exposed individuals that presented the outcome after a given amount of time (Fig. 8.2). In oncology, cohort studies are mostly used to study risk factors or prognostic factors (more details below).

8.3.1.1 Examples

Cohort Study 1: A cohort of male British doctors [9]. This was a prospective cohort study conducted with British physicians as the target population. The objective of the study was to evaluate the association of tobacco consumption with the occurrence of diseases and mortality in the long term. In 1951, questionnaires about smoking habits were sent to 34,440 doctors from the British Medical Association. Follow-up of the participants began when doctors first answered the questionnaire. Subsequent follow-up was done at 13, 20, 40, and 50 years later. Smoking habits were then associated with mortality and its causes (including cancer) [9].

Cohort Study 2: A cohort of patients with head and neck squamous cell carcinoma (HNSCC) [10]. This was a prospective cohort study having patients with HNSCC in Brazil as the target population. The objective of the study was to evaluate the association of the presence of HPV with cancer mortality. A total of 1093 patients were recruited at the time of diagnosis. Epidemiological questionnaires were employed, and biological samples (tissue and blood) were collected. Recruitment occurred between 1998 and 2008. Of the 1093 patients, HPV 16 serology and HPV16 DNA detection were performed in blood and tumor tissue from 398 patients [10].

Cohort Study 3: A cohort of patients with osteosarcoma [11]; this study had patients with osteosarcoma as the target population. The objective of the study was to evaluate the association of HULC gene expression with the clinical outcome of the patients (recurrence, progression, or death due to cancer). In the period from 2006 to 2013, 175 patients were diagnosed with osteosarcoma, from whom the relevant cryopreserved biological material was available (blood and tissue). This was a retrospective study and, at the time of data acquisition, the patients had already been diagnosed and biological material had been stored. Moreover, outcomes had already occurred; patients' charts were analyzed for patient selection and data collection. However, after verification of the inclusion criteria, only 76 cases were selected for the study. The quality of the stored biological material was also verified and, from these 76 cases, only 33 samples were adequate for RNA extraction and, consequently, for the assessment of HULC expression [11].

Table 8.5 presents a summary of the causes, times, and outcomes examined in these example cohort studies.

8.3.1.2 Exposure

Exposure refers to any feature that is believed to influence the occurrence of the outcome. Exposure can be interpreted as a risk factor or a prognostic factor, depending on the purpose and design of the study.

Table 8.5 Summary of cause, time, and outcome of Cohort Studies 1, 2, and 3

	Cause	Minimum follow-up time	Outcome
Cohort study 1	Smoking	15; 20; 40, and 50 years	Disease occurrence/death
Cohort study 2	HPV16	>6 months	Death
Cohort study 3	HULC	>13 months	Recurrence, progression, or death

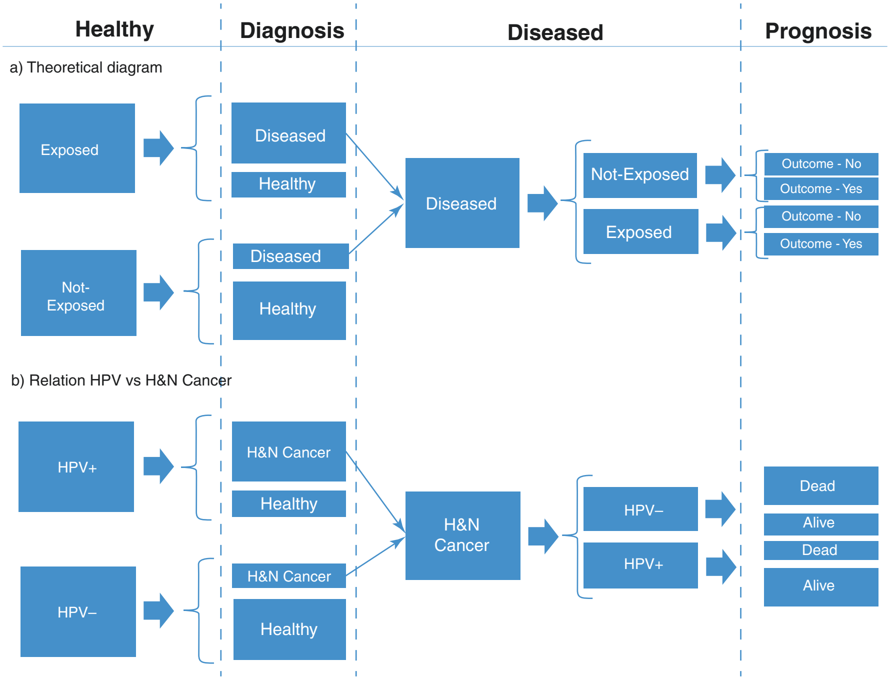


Fig. 8.3 (**a**) Diagram of the theoretical association between risk factors and prognostic factors; (**b**) association between human papillomavirus (HPV) and head and neck squamous cell carcinoma (HNSCC) as a risk factor and as a good prognostic factor

- **Risk factor**: a characteristic (or attribute) of a subgroup of the population with higher disease incidence (outcome) compared with the group that does not present this characteristic.
- **Prognostic factor**: a characteristic (or attribute) that can be associated with the course of the disease (outcome). We usually define as good prognostic factor characteristics those related to a better patient outcome.

Cohort Study 1 above is an example of a study that aimed to verify the risk factors, whereas in Cohort Studies 2 and 3 the aim was to study the prognostic factors.

In Fig. 8.3a, we present the relationship between risk and prognostic factors for a disease under study. Risk factors are associated with characteristics that influence the onset of the disease; thus, at the beginning of follow-up, all elements of the population are healthy. The population is then divided into subgroups according to the characteristics of interest and according to what is believed to be associated with the incidence of the disease. In this case, the disease is considered the outcome. In studies of prognostic factors, we evaluate individuals that present the disease of interest at the initial time; thus, healthy individuals are not followed, only patients. In those patients, we assess the characteristics that are influencing the course of the disease. Thus, there is a change in the characteristics of the population at risk of developing a disease when compared with the population used for studying the prognosis of the disease.

As there is a change in the study population, a characteristic that is a risk factor for the occurrence of the disease might not, necessarily, be a prognostic factor, or vice versa. However, there are situations in which a characteristic that is linked to a risk of disease occurrence is also a prognostic factor concerning the course of this disease; for example, the relationship between HPV and HNSCC in the study reported in [10] (Fig. 8.3b).

It is known that the presence of HPV increases the risk of HNSCC development [12]. However, in the Cohort Study 2 example, once patients presented with HNSCC, those with HPV-negative tumors had a worse prognosis (death due to cancer) than those with HPV-positive tumors [10]. In other words, HPV is a risk factor for HNSCC, but an HPV-induced tumor presents a better prognosis.

8.3.1.3 Time

Time is another important component in a cohort study and should be planned very carefully. We will explain this topic by answering the following questions:

- How long should we follow the cohort? (follow-up time).
- How do I follow-up? (data collection).
- How do I measure the outcome over time? (ways to measure the outcome).

Follow-Up Time

When planning the study, we need to understand the behavior of the natural course of the disease, as this will help define for how long the cohort should be followed [13]. Thus, in oncological studies in which the outcome is death (prognostic factor studies), it is common to include a follow-up period of 5 years, such as in a breast cancer population. A follow-up of 2 years may not be reasonable in most types of cancers, because that is too short a time for evaluating death. In populations in which the disease is aggressive, such as small cell lung cancer or pancreatic cancer, in which the vast majority of deaths due to cancer occur within 2 or 3 years, a follow-up time of 5 years might not be necessary, as perhaps 2 years of follow-up is sufficient to observe and analyze this outcome. In thyroid cancer, on the other hand, in which the prognosis is very good and the occurrence of death is delayed, a follow-up period of more than 10 or 15 years might be necessary for survival analysis. An important detail is that, ideally, all individuals in the study population should be followed for the proposed period or until the outcome occurs. In cohort studies, sometimes the relevant follow-up information is not available for all patients (those considered as being lost to follow-up). Despite the fact that statistical models, such as Kaplan-Meier, can deal with incomplete follow-up information, this 'lost to follow-up' information can introduce an important bias to the analysis and results, as discussed in more detail below.

In Cohort Study 1, we observed follow-up times of up to 50 years. In the first publication [9], despite the original proposal being to study the effect of smoking on mortality, 13 years was not a sufficient time for verifying the effect of tobacco; however, it was enough to consider the effects of exposure to alcohol. Therefore, the researchers then evaluated the influence of alcohol consumption on mortality. In other studies, with 20, 40, and 50 years of follow-up, it was possible to evaluate tobacco exposure in regard to its association with mortality.

Data Collection

It is important to define how to collect the data. For example, will the researcher follow the participant from the time of inclusion, or try to retrieve the information retrospectively? Thus, there are two possibilities for considering follow-up and data collection:

- **Prospective Cohort Study**: The exposure is verified in the present, the researcher defines a sampling of the population of interest that does not present the outcome at the moment, and may divide the sample into groups according to exposure (exposed and not-exposed); the researcher follows the participants for the determined period of time or until they present the outcome of interest (future), and later conducts the analyses. Thus, the follow-up of the patients is prospective (Fig. 8.4a).
- **Retrospective Cohort Study**: We observe (at present) whether or not individuals have presented the outcome, while exposure (or not) occurred in the past. Subsequently, the researcher collects exposure information from the past, and analyses are performed. Thus, follow-up occurs retrospectively (Fig. 8.4b).

IMPORTANT: We must be careful not to confuse a retrospective cohort study with a case-control study (Chap. 9). When selecting the sample for a cohort study, we do not control how many participants in the groups have (or not) the outcome, whereas in a case-control study, the researcher defines the participants in each group based on the presence of the disease or outcome.

Cohort Study 2 is a prospective cohort study, as patients entered the study at the time of diagnosis and were followed-up afterward. In Cohort Study 3, patients had already been diagnosed and treated at the time of data collection. The researcher started data collection at least 13 months after diagnosis, which makes this study a retrospective cohort study.

A major problem encountered in cohorts is the access to and the quality of data. In a prospective cohort, it is often not possible to follow all patients over time, losing some of them (and thus, their information) during the follow-up. In a retrospective cohort, there is a need to seek information from secondary sources (such as medical records) or the memory of the patient. Such sources of information are not always reliable or complete, thus introducing bias in the study.

In Cohort Study 3, a major problem that the researchers faced (one that is very common in studies using stored biological samples) was the degradation of the biological material. All patients had provided biological samples, but the researchers found that these were not adequate for RNA extraction at the time of the study and,

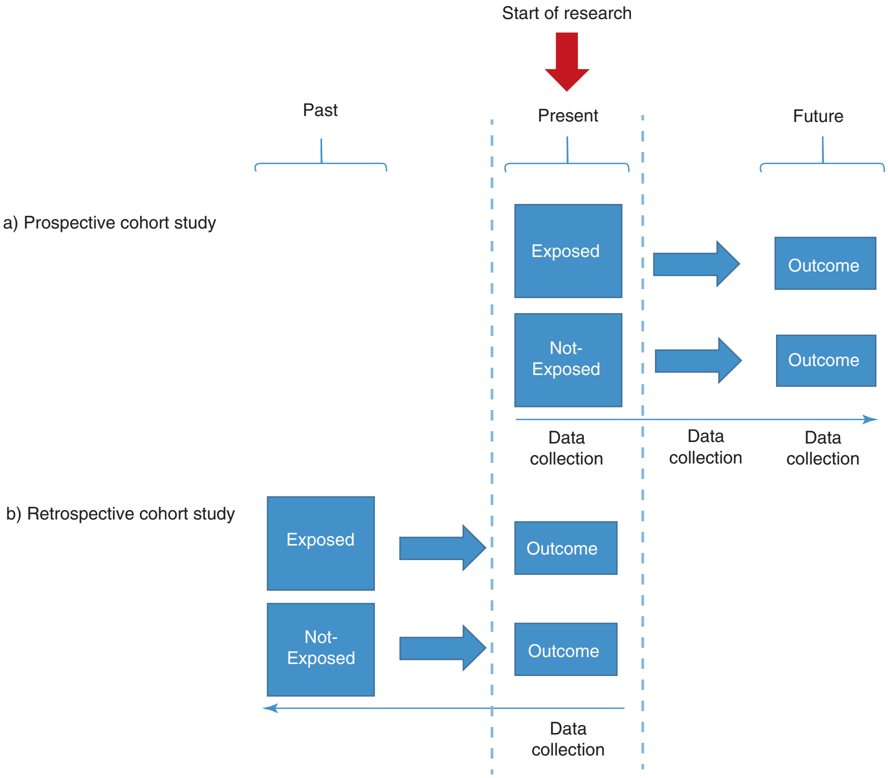


Fig. 8.4 Theoretical scheme representing the type of study follow-up. (**a**) Prospective cohort study; (**b**) retrospective cohort study

consequently, the expression of the HULC marker could not be measured, which culminated in a drastic decrease in sample size. From 175 possible cases, only 76 were selected after employing the inclusion criteria of the study (for example, patients who had already started treatment at the time of collection of biological material). This number was further decreased to 33 owing to biological sample degradation. As such, only 18.8% of the cases originally available were ultimately analyzed. This degradation of biological samples is an important fact to be considered when planning to use such samples for molecular studies with currently available technologies.

8.3.2 Ways to Measure the Outcome

When patients are followed, there are two options for measuring the outcome. We must plan how the collection will be conducted depending on whether we choose one-measure or repeated-measures for the outcome:

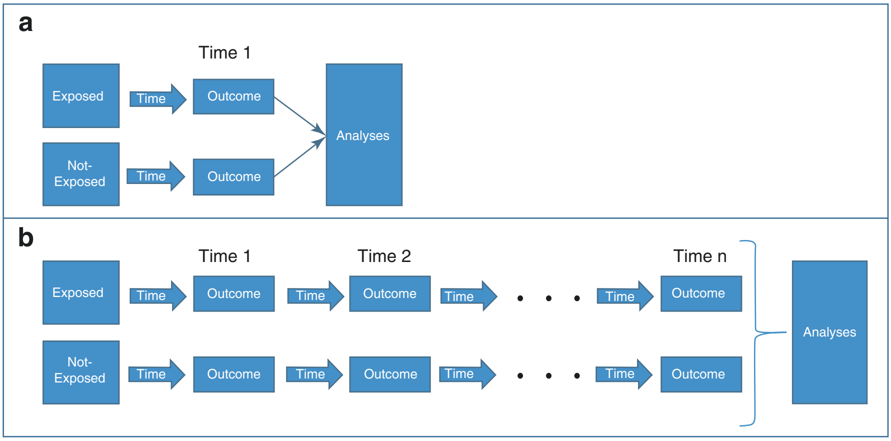


Fig. 8.5 Theoretical scheme representing the collection mode: (**a**) One measure; (**b**) repeated measures

- **One-measure**: In this type of study, the outcome in study participants is measured at one point in time, and then the statistical analyses are performed, associating exposure and outcome (Fig. 8.5a). This measure is mainly used for survival analysis.
- **Repeated-measures**: In this type of study, the study participants are monitored over time, and the outcome is measured at predefined times (for example, one month after surgery, two months after surgery, etc.). The time-points must be standardized, and the same outcome is measured at the different time-points. In this analysis, measurements are taken into account at all times, and not only at the first or last time-points (Fig. 8.5b). Repeated measures are used, for example, in studies evaluating trends in individuals' quality of life or other patient-reported outcomes over time.

Cohort Study 1 reported results after 13, 20, 40, and 50 years of follow-up. At first, it may appear that this is a study with repeated measures. However, the study showed updates of the outcome. First, researchers divided the group according to exposure; 13 years passed and the outcome was evaluated at a single time point. Subsequently, in 20 years, mortality was updated, and data were re-analyzed using the updated information, and so on, which is characteristic of a study with one measure.

8.3.2.1 Outcome

Cohort studies in oncology usually include the occurrence of cancer or disease progression as the main outcomes. However, the concept of outcome is much broader and may refer to disease occurrence, exacerbation of illness, and score measurements, among other factors. Regardless of the type of study, an outcome can be classified as binary, continuous, or time-to-event [14].

- Binary: The outcome is presented in two categories, representing presence or absence of a given attribute. For example, diseased or healthy, dead or alive.
- Continuous: The outcome is presented with a numeric variable (discrete or continuous). For example, quality-of-life scores, prostate-specific antigen (PSA) levels (quantitative), and expression of tumor markers.
- Time-to-event: The outcome is presented as the time until the occurrence of an event. The event is usually dichotomous and is characterized by exacerbation of the disease; for example, cancer-specific survival (time to death due to cancer).

In Cohort Study 1, the outcome was death (and its cause); the study had a binary outcome (alive vs. death from disease). In Cohort Study 2 and Cohort Study 3, the outcome was considered as time-to-event. Cohort Study 2 considered the time to the occurrence of metastasis during treatment or death, whereas in Cohort Study 3, the outcome was defined as time to death due to any causes (usually known as overall survival).

When we work with a time-to-event outcome, our interest is not only the outcome but also the time to the occurrence of this outcome; thus, we use techniques for survival analysis (Chap. 7). In oncology, specific terms are used for survival; for example, overall survival, disease-free survival, and progression-free survival, among others.

Of note, Punt et al. [15] published a review identifying key endpoints in studies with survival analysis in colon cancer. The review showed that, although different articles referred to survival using the same term, there was no standardization of the definition of the event and censoring. Among other findings, the authors reported that, in a sample of 44 papers estimating disease-free survival, 13 papers (29.6%) defined the event of interest (endpoint) as recurrence, second primary tumor, or death; 10 (22.7%) defined that event as recurrence or death; 4 (9.1%) defined it only as recurrence; and 17 (38.6%) did not indicate what was considered to be the endpoint. Thus, although different studies use the same terms for survival, many consider different possible outcomes, which makes comparisons with other studies inadequate. In the review, the authors strongly suggested a standardization method to define events and censoring for each survival type.

8.3.2.2 Relative Risk

According to the possible combinations of exposure, time, and outcome, the statistical analysis of the data may become very complex, and this requires statistical assistance. However, a cohort study allows us to calculate the incidence of the disease among exposures, which in turn, permits the calculation of the relative risk [16]. This is the *only* type of study, among observational studies, which has this characteristic.

Incidence: The proportion of individuals presenting the disease (outcome) among individuals at a given exposure.

Table 8.6 Relationships between exposure and outcome, and relative risk (RR) in a cohort study

	Outcome		
Exposure	**Yes**	**No**	
Yes	Individuals exposed with outcome	Individuals exposed without outcome	
No	Individuals not exposed with outcome	Individuals not exposed without outcome	
	Disease	**No Disease**	**Total**
Exposed	**a**	**b**	**a + b**
Not-exposed	**c**	**d**	**c + d**
Total	**a + c**	**b + d**	**N**
Example:			
Frequencies	**Lung cancer**	**Healthy**	
Smokers	**133**	**102,467**	**102,600**
Non-smokers	**3**	**42,797**	**42,800**
Total	**136**	**145,264**	**145,400**

RR = [133/(102,600)]/[3/(42,800)] = 18.49
aRR = $[a/(a + b)]/[c/(c + d)]$RR = [133/(102,600)]/[3/(42,800)] = 18.49

Relative risk: The relative risk (RR) is the ratio between the incidence of the disease (outcome) among exposed individuals and that among unexposed individuals.

Suppose that the exposure divides the population into two groups (two possibilities) and that the outcome is of binary type. We present our sample in a 2 × 2 cross-reference table, in which the rows describe the exposure group and the columns indicate the outcome, as depicted in Table 8.6.

We can define the incidence as:

- $a/(a + b)$: proportion of individuals presenting the outcome among the exposed group.
- $c/(c + d)$: proportion of individuals presenting the outcome among the not-exposed group.

  And we can define RR as:
- $[a/(a + b)]/[c/(c + d)]$: the ratio between the incidence of the outcome among those exposed and the incidence of the outcome among not-exposed individuals.

The RR indicates how much more likely it is that an exposed individual will develop the outcome. Thus, in the example in Cohort Study 1, we can infer that smokers had 18.49 times the risk of developing lung cancer compared with non-smokers.

In general, we can interpret RR as:

- RR < 1: Protective risk or prognostic factor, as it reduces the chance of exposed individuals to have the outcome in comparison with not-exposed individuals.

- RR = 1: The exposure cannot be considered a risk or prognostic factor, as it does not increase or decrease the chance of the exposed group presenting the outcome when compared with the not-exposed group.
- RR > 1: The exposure is considered a risk or prognostic factor, as it increases the chance of the exposed individuals presenting the outcome when compared with the not-exposed group.

In the literature some articles report using the odds ratio (OR) to measure associations in exposed and not-exposed groups in cohort studies; this is particularly likely for retrospective studies. Although it is somewhat controversial to use the OR in cohort studies, the interpretation of the result should be similar to the RR. In cohort studies in which the outcome occurs in less than 10% of the unexposed population, the OR provides a reasonable approximation of the RR. However, when an outcome is more common, the OR will overstate the RR [17].

8.3.3 Lost to Follow-Up

The purpose of a cohort study is to monitor individuals for a determined time and verify which individuals present the outcome. However, during follow-up, we can lose information about patient outcome for various reasons, and this may interfere with the analysis results. These cases are known as lost to follow-up. Little and Rubin [18] classify the occurrence of lost to follow-up into three different mechanisms:

- Missing completely at random (MCAR): missing cases occur randomly, with no relation to the exposure or outcome.
- Missing at random (MAR): missing cases are related to some exposures, but not to the outcome.
- Not missing at random (NMAR): missing cases are related to the outcome, whether owing to the exposure or not.

8.3.3.1 Example: MAR

In Cohort Study 3, there were originally 175 samples, but this number decreased to 33 owing to sample degradation. Degradation occurs because of technical problems (collection or storage) or storage time. Thus, it can be assumed that the degradation of the material is not influencing the outcome (event), but rather only preventing us from obtaining relevant information about the patients (HULC expression). Thus, this is an example of MAR.

8.3.3.2 Example: NMAR

In a study measuring the quality of life in cancer patients over a long period of time, it is to be expected that patients with a poor prognosis will die before the end of the follow-up period. Thus, such a study would not predefine all the data collection periods. In such studies, the exclusion of patients with an incomplete follow-up would decrease the statistical power related to sample size, as well as introducing bias in the results, as the quality of life of patients who died would have been worse than that of those who were alive by the end of the follow-up period. As such, we would be overestimating the quality-of-life indexes, a feature of NMAR.

In Cohort Study 1, the results are based only on respondents participating in the survey. What causes an individual to not respond to a survey may be related to good health or lack thereof. Thus, we have to assume that the chance of someone not responding is not random (NMAR).

8.3.3.3 What Percentage of Lost to Follow-Up Is Acceptable in a Study?

Kristman et al. [19] simulated various scenarios considering the three mechanisms of lost to follow-up for different loss percentages (0%, 5%, 10%, 30%, 40%, 50%, and 60%), and each scenario was replicated 1000 times. They showed that for the MCAR and MAR mechanisms there was no significant difference between the estimates for the different percentages of lost to follow-up, with the mechanisms not biasing the results and only affecting the test power (see Chap. 5). But for the NMAR mechanism, as we increase the proportion of lost to follow-up, we increase the error, introducing bias in the estimate (in this case, by decreasing the RR). For the MCAR and MAR mechanisms for all loss percentages, the coverage (percentage of confidence intervals including the true RR) was around 95%. As for the NMAR mechanism, the probability of coverage (PC) had a significant drop after 10% of lost to follow-up, as shown in Table 8.7.

Thus, we can say that the MCAR and MAR mechanisms are non-informative, as their loss does not entail bias for data analysis. On the other hand, the NMAR mechanism is an informative loss, since the loss of its information is associated with the outcome, and it is necessary to verify the possibility of using statistical methods for the imputation.

The findings of Kristman et al. [19] support what Simon and Wittes [20] described in an editorial on guides for clinical trial studies. In this guide, they reported that up to 15% of incomplete information is "acceptable," and that when that rate exceeds 20%, it indicates inadequate selection of patients or inadequate planning of the study. However, in their editorial, Simon and Wittes do not justify the reasoning behind the use of these particular percentages.

Table 8.7 Probability of coverage (PC) for MCAR, MAR, and NMAR mechanisms, varying the percentage of lost to follow-up

Percentage of lost to follow-up	5%	10%	20%	30%	40%	50%	60%
PC-MCAR	97%	95%	95%	95%	95%	96%	95%
PC-MAR	95%	95%	95%	95%	95%	95%	95%
PC-NMAR	95%	93%	83%	70%	39%	18%	5%

MCAR Missing completely at random; *MAR* missing at random; *NMAR* not missing at random

Cohort studies in which the interest is the time to occurrence of a particular outcome use a survival analysis approach, in which incomplete data information is introduced with the concept of censoring. These cases are kept in the analyses and the date of the last objective information from the patient (probably the date of the last visit) is indicated at the end of follow-up, censoring the case. Priante et al. [21] presented the effect of lost to follow-up on a cohort of patients with head and neck cancer. This study initially estimated the survival considering the follow-up time according to the last visit of the patient presented in the medical charts (52.3% initial rate of lost to follow-up). Subsequently, an active search was made to update the vital status of all patients (the lost to follow-up rate then dropped to 17.3%), and a revised Kaplan-Meier estimate was calculated. When comparing the two estimates, the authors found that the survivals at 5 and 10 years for the first situation (52.3% lost to follow-up) were 54.0% and 46.0%, respectively, whereas for the second situation (17.3% lost to follow-up) estimates were 42.8% and 28.2%, respectively, indicating a significant difference between the estimates, due exclusively to censoring many patients based on incomplete information on vital status, and mostly probably denoting an NMAR mechanism.

A difficult problem is to identify which of the three situations corresponds to incomplete data in a study. One possibility is the comparison of clinical and demographic characteristics between the groups with complete vs. non-complete data. If the characteristics are statistically similar, it is probably a not-informative case (MCAR or MAR). If there is a difference between the groups, it is an informative case (NMAR). Unfortunately, it is expected that the MCAR or MAR mechanisms would occur less frequently, since it is expected that the loss of information is usually associated with the exposure of interest or the outcome. As NMAR is much more frequent and it does introduce relevant bias, it is suggested that NMAR should always be considered. In fact, the literature suggests that NMAR is the most likely mechanism for follow-up cohort studies, because subjects who drop out tend to have different outcomes from those who remain in a study and, for this reason, minimizing the chance of lost to follow-up is very important for internal and external validation of the study results [19, 21].

IMPORTANT: Regardless of the type of incomplete data, lost to follow-up cases should never be excluded, as this exclusion could cause an even worse bias in the analysis results. Researchers should, rather, use adequate statistical resources that incorporate incomplete information into their calculations.

8.3.3.4 Lost to Follow-Up in One Measure

In this situation, we can find the loss either in the outcome or in the exposure. Cohort Studies 1 and 3 characterize this example. We can only analyze patients who have complete information (those who responded to the survey or have biological material for molecular analysis) or use imputation methods to estimate the missing values.

8.3.3.5 Lost to Follow-Up in Studies with Repeated Measures

In studies with repeated measures, there are two possibilities for loss, related to monotonic and non-monotonic outcomes [14]:

- Monotonic: When, after the first loss, we get no further measurements from that individual.
- Non-monotonic: There is no pattern in the occurrence of the loss. For example, in a study with many time points, we can lose the patient at time point 2 but he/she returns at other time-points of the study.

In both cases, we can use imputation methods or use a statistical methodology to analyze only the variables in the study. But one would need to obtain the exposure information and the outcome measurement for at least one time point. Otherwise, the participant will be excluded (complete loss).

8.3.4 Advantages and Disadvantages of Cohort Studies

Cohort studies are considered as one of the best types of observational studies in regard to their level of scientific evidence, second only to clinical trials (Chaps. 10, 11, 12, and 13). However, these studies have advantages and disadvantages. We now list the main strengths and weaknesses of this type of study.

8.3.4.1 Advantages

Ethical issues: Clinical trials are often unfeasible owing to ethical problems (for example, we cannot conduct a clinical trial where we "force" a group of people to smoke to investigate the effect of smoking on mortality). Cohort studies are a valuable alternative, with a relevant level of scientific evidence.

Cause-effect relation: As there is monitoring over time, and at the beginning of the follow-up patients do not present the outcome, we can clearly define the cause-effect relationship (causality).

Incidence coefficient: As it is possible to define a cause-effect relationship, we can calculate the incidence coefficient of the outcome between exposed and not-exposed individuals and, consequently, obtain important epidemiological estimates.

Natural history of the disease: As we follow the patients from exposure to outcome, we can study the natural history of the disease.

Quality in data collection: In a prospective cohort study, since the data collection is prospective, data can be collected with better quality and with less bias, as we can measure information at the exact time it is collected.

Multiple exposures: Cohort studies allow multiple exposures to be evaluated simultaneously. For example, one can check multiple tumor markers or patients' characteristics and habits.

Multiple outcomes: Cohort studies allow several outcomes to be evaluated in the same study. For example, in a single study, we can analyze both overall survival and disease-free survival.

8.3.4.2 Disadvantages

Rare diseases: Cohort studies are not suitable for studies in which the outcome occurs infrequently (for example, rare diseases), since we would need to include too many participants to quantitatively investigate an outcome, which is often unfeasible.

Follow-up time: Depending on the natural history of the disease, it may take too much time to observe the outcomes (this problem is worse in prospective cohorts).

Lost to follow-up: As cohort studies may have long follow-up periods, it is usual to lose track of some study participants throughout the process.

High cost: Prospective cohort studies are expensive, as they require patient follow-up.

Confusion factors: In a retrospective cohort study, as we collect information retrospectively, we are unable to adequately measure information on confounding factors.

Conclusion

Cross-sectional and cohort studies are important epidemiological study designs in oncology. They have very distinct and well-defined characteristics and allow for different measurements.

Cross-sectional studies measure prevalence at a low cost, are easy to conduct, and allow for the rapid collection of data and real-time information, with no need to follow the individuals, and they work well to monitor population health and for planning in public health. However, this study design does not measure temporality, nor does it allow for determining causality; it is not fit for low-prevalence outcomes and is subject to bias in selection and information, both these factors influencing its internal and external validity.

Cohort studies are considered to be the best type of observational studies in regard to their level of scientific evidence, second only to clinical trials. They are designed to study the natural history of the disease and to measure causality, which means that there is a need to follow the individuals for a determined period to observe the possible occurrence of the outcome. This type of study design allows the measurement of association for risk or prognostic factors and also allows several outcomes to be evaluated in the same study. However, the study design also permits bias in lost to follow-up data and information; this is particularly frequent in retrospective cohort studies when information bias and adequate measurement of confounding factors are a problem. This study design is not suitable for studies in which the outcome occurs infrequently (rare diseases or very high survival rates).

References

1. de Sanjose S, Quint WG, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. Lancet Oncol. 2010;11(11):1048–56.
2. Costa RFA, Longatto-Filho A, Pinheiro C, Zeferino LC, Fregnani JH. Historical analysis of the Brazilian cervical cancer screening program from 2006 to 2013: a time for reflection. PLoS One. 2015;10(9):e0138945.
3. Zlotta AR, Egawa S, Pushkar D, Govorov A, Kimura T, Kido M, et al. Prevalence of prostate cancer on autopsy: cross-sectional study on unscreened Caucasian and Asian men. J Natl Cancer Inst. 2013;105(14):1050–8.
4. Rogentine GN Jr, Yankee R, Gart J, Nam J, Trapani R. HL-A antigens and disease: acute lymphocytic leukemia. J Clin Investig. 1972;51(9):2420.

5. Nalbandian A, Yan B, Pichugin A, Bronson R, Kramnik I. Lung carcinogenesis induced by chronic tuberculosis infection: the experimental model and genetic control. Oncogene. 2009;28(17):1928–38.
6. Libshitz HI, Pannu HK, Elting LS, Cooksley CD. Tuberculosis in cancer patients: an update. J Thorac Imaging. 1997;12(1):41–6.
7. Lin W-W, Karin M. A cytokine-mediated link between innate immunity, inflammation, and cancer. J Clin Invest. 2007;117(5):1175–83.
8. Silva DR, Valentini DF Jr, Müller AM, de Almeida CP, Dalcin Pde T. Pulmonary tuberculosis and lung cancer: simultaneous and sequential occurrence. J Bras Pneumol. 2013;39(4):484–9.
9. Doll R, Peto R, Hall E, Wheatley K, Gray R. Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. BMJ. 1994;309(6959):911.
10. López RVM, Levi JE, Eluf-Neto J, Koifman RJ, Koifman S, Curado MP, et al. Human papillomavirus (HPV) 16 and the prognosis of head and neck cancer in a geographical region with a low prevalence of HPV infection. Cancer Causes Control. 2014;25(4):461–71.
11. Uzan VRM, van Helvoort Lengert A, Boldrini É, Penna V, Scapulatempo-Neto C, Scrideli CA, et al. High expression of HULC is associated with poor prognosis in osteosarcoma patients. PLoS One. 2016;11(6):e0156774.
12. Mork J, Lie AK, Glattre E, Clark S, Hallmans G, Jellum E, et al. Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. N Engl J Med. 2001;344(15):1125–31.
13. Compton CC, Byrd DR, Garcia-Aguilar J, Kurtzman SH, Olawaiye A, Washington MK. AJCC cancer staging atlas: a companion to the seventh editions of the AJCC Cancer staging manual and handbook. New York, NY: Springer; 2012.
14. Srivastava DK, Robison LL, Wu X, Rai SN. Design and analysis of cohort studies: issues and practices. Biom Biostat Int J. 2015;2(5):1–7.
15. Punt CJ, Buyse M, Köhne C-H, Hohenberger P, Labianca R, Schmoll HJ, et al. Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. J Natl Cancer Inst. 2007;99(13):998–1003.
16. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst. 1951;11(6):1269–75.
17. Viera AJ. Odds ratios and risk ratios: what's the difference and why does it matter? South Med J. 2008;101(7):730–4.
18. Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 2014.
19. Kristman V, Manno M, Côté P. Loss to follow-up in cohort studies: how much is too much? Eur J Epidemiol. 2004;19(8):751–60.
20. Simon R, Wittes RE. Methodologic guidelines for reports of clinical trials. Cancer Treat Rep. 1985;69(1):1–3.
21. Priante AVM, Carvalho AL, Ribeiro KCB, Contesini H, Kowalski LP. The importance of long-term follow-up of head and neck cancer patients for reliable survival analysis. Otolaryngol Head Neck Surg. 2005;133(6):877–81.

Design of Retrospective and Case-Control Studies in Oncology

9

Katherine S. Panageas, Debra A. Goldman, and T. Peter Kingham

9.1 Introduction

The objective of research studies is to make inferences about hypothesized relationships within a population. These relationships include differences in survival among treatment groups, various risk factors for surgical outcomes, differences in quality of life, and genetic variations among cancer subtypes. The study design used to answer the research question is critical for the ability to draw conclusions and is directly related to the statistical analysis methods that can be applied. Properly designed and executed studies provide the strongest level of empirical evidence.

9.1.1 Randomized Controlled Trials

The gold standard study design for clinical research is the randomized controlled trial (RCT), which is the most likely to minimize inherent biases. In RCTs, using a large enough sample size, randomization ensures that each patient has an equal chance of receiving a given treatment and that treatment groups are comparable with respect to any known or unknown factors that may affect the outcomes. In addition to eliminating selection bias, randomization provides a simple foundation for straightforward statistical analyses compared with observational studies. Despite being considered the gold standard, RCTs have several drawbacks. First, they are expensive and time-consuming, and they require organizational infrastructure to

K. S. Panageas, Dr.P.H. • D. A. Goldman, M.S.
Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center,
New York, NY, USA

T. P. Kingham, M.D. (✉)
Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: kinghamt@mskcc.org

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_9

develop and conduct. Second, RCTs for infrequent or rare outcomes require sizable sample sizes, so these outcomes may be more practical to examine using alternative study designs. Third, RCTs may pose ethical problems or may not be feasible owing to difficulties with recruitment and compliance. Fourth, results from RCTs may not be generalizable to real-world populations or circumstances, in which the environment cannot be strictly controlled [1, 2].

9.1.2 Observational Studies

Observational studies are alternatives to RCTs, and can be either prospective or retrospective. Observational studies may be used in settings when it is unethical to randomize patients to receive specific treatments, or to provide preliminary evidence for hypotheses of RCTs.

9.1.2.1 Prospective Observational Studies

In a prospective observational study, data collection and the events of interest occur in a group of individuals, some of whom have had, currently have, or will have the exposure of interest, such as a certain treatment, to determine the association between that exposure and the outcome. However, prospective observational studies are limited to conditions that occur relatively frequently and to studies with relatively short follow-up periods, so that sufficient numbers of eligible individuals can be enrolled and followed within a reasonable study period.

9.1.2.2 Retrospective Observational Studies

All retrospective research studies are classified as observational studies because the allocation to treatment or assignment of factors is not under control of the investigator. In retrospective studies, the study sample is generated from secondary or preexisting data. The disease experience of the group between a defined time in the past and the present is then reconstructed from medical records. Compared with prospective studies, retrospective studies are inexpensive, as they make use of available information. Further, retrospective studies of rare conditions are much more efficient because individuals experiencing these rare outcomes can be found among patient records rather than the investigators needing to prospectively follow a large number of individuals to identify a few cases. Studies have shown that the majority of publications in clinical subspecialty journals are based on retrospective observational studies [3–5]. Also, as most medical centers transition from paper to electronic medical records and as computing power advances to handle ever larger data sets, retrospective studies are becoming easier and more efficient to conduct.

Retrospective studies have long-established use in surgical oncology [6–8]. Single-institution data or large multicenter efforts examining past experiences can serve many beneficial purposes, including generating hypotheses to develop future prospective studies, to explore ideas in translational laboratory research projects, or to compare results with previous studies that enrolled a smaller or more heterogeneous patient population. Further, retrospective analyses can provide critically relevant data for populations known to be poorly represented in clinical trials—especially of cancer—including older adults and individuals with eligibility-restricting comorbidities. These analyses also may identify adverse events that are potentially unrecognized in the often highly homogenous groups of study participants. Finally, both the safety and the efficacy of treatment afforded by longer observation periods and more prolonged therapy can be revealed by retrospectively examining previously treated patients [9]. However, because retrospective studies do not involve randomization, the potential for significant biases exists, such as sample selection and recall and referral biases, which can limit the applicability and generalizability of these studies.

9.2 Types of Retrospective Studies

The historical cohort study and the case-control study are two of the most common retrospective designs. A retrospective **cohort study** comprises a sample of individuals (e.g., surgically resected pancreatic cancer patients) in whom we assess the relationship between risk factors and outcomes, such as post-surgical complication rates, disease recurrence, and overall survival. Risk factors are considered the **exposure**, a broad term used to denote any factor that is potentially related to the outcome of interest [10]. In contrast, in a retrospective **case-control study**, the outcome (e.g., post-operative complications) is measured *before* the exposure. Controls are selected from a pool of patients who have not experienced the outcome (Fig. 9.1). It is critical that the control group be as similar to the cases as possible in terms of other factors, such as demographic and treatment details [11, 12]. The retrospective case-control study is an important research strategy encountered in the medical

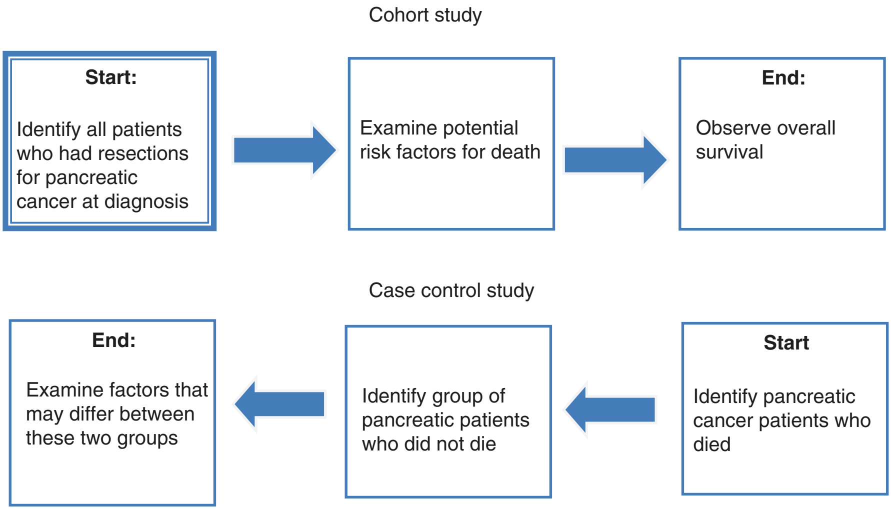

Cohort study
Start:
Identify all patients who had resections for pancreatic cancer at diagnosis
Examine potential risk factors for death
End:
Observe overall survival
Case control study
End:
Examine factors that may differ between these two groups
Identify group of pancreatic patients who did not die
Start
Identify pancreatic cancer patients who died

Fig. 9.1 Illustration of the differences between cohort and case-control studies

literature, and if carefully executed, can be an invaluable source of clinical information. Unfortunately, the retrospective viewpoint of case-control studies—looking "backwards" from an outcome event to an earlier exposure—is accompanied by numerous methodological hazards, including recall bias, which will be discussed later in this chapter.

Retrospective studies are often criticized on methodological grounds. Researchers must pay careful attention to selecting appropriate study groups, defining and detecting the outcome event, defining and ascertaining the exposure, assuring that the compared groups were equally susceptible to the outcome event at baseline, and performing careful statistical analysis. If systematic bias enters the research at any of these points, erroneous conclusions can result. In this chapter, we will cover design topics specific to retrospective studies, including validity, confounding, sample selection, sampling methods, missing data, and considerations for particular oncology outcomes.

9.3 Validity

The quality of a study depends on many factors, including internal and external validity. **Validity** is the degree to which a study result is likely to be true and free from bias [13]**.** As mentioned, **retrospective** study designs are inherently more susceptible to bias, given the lack of control over group assignment and the experiment environment. The study design and execution greatly determine the **internal validity. A study is internally valid if reported differences can be attributed to the exposure or intervention** [14] **and cannot be attributed to selection bias, information bias, or confounding. Confounding** is the distortion of the effect of one risk factor by the presence of another (Fig. 9.2). In randomized studies, confounding is typically accounted for in the randomization process. In retrospective studies, confounding can be controlled by restriction sampling, by matching on the confounding variable, or by accounting for it in the analysis using multivariable modeling.

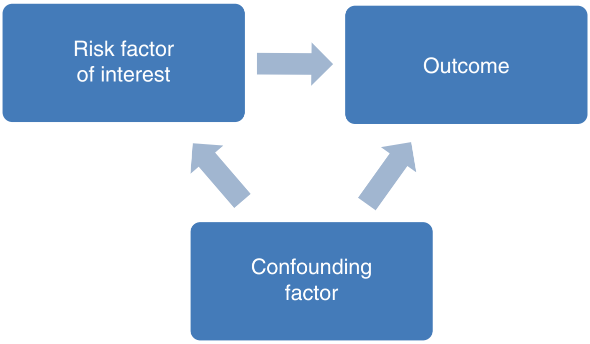


Fig. 9.2 Illustration of confounding

If a study is internally valid, it is important to assess whether it is has external validity as well. **External validity** refers to the extent to which the results can be generalized to other populations, other settings, and across time [15, 16]. For instance, if we build a model to predict survival in patients with incidental gallbladder cancer, we would want that model to be predictive for patients at other centers, in future years, and other circumstances. External validity is highly related to applying the appropriate sampling techniques, as we will now explore.

9.4 Sampling

Sampling refers to the process of selecting individuals to be included in a study. A **representative sample** is one in which the group sufficiently embodies the population that one is attempting to study, known as the **target population**. In retrospective samples, representativeness, generalizability, and sampling issues are important considerations. Unlike prospective research, in which one can control who is evaluated through enrollment and eligibility criteria, and one can control treatment environment and outcomes assessments, in retrospective research, one is limited by external factors that may have affected who is included in the study sample and who is not.

For example, in a study evaluating outcomes for gallbladder cancer over time, sampling may be limited to a single research institution. If the institution is a referral center for more complex cases or more advanced stage patients, the sample may not represent gallbladder cancer outcomes at other institutions or gallbladder cancer patients as a whole. Additionally, if data were retrieved from an institutional surgical database, sampling would be restricted to those patients who received consultation from a surgeon. Patients seen only by a medical oncologist would not be included, so findings could not be generalized to all patients with gallbladder cancer, but rather only to those who received surgery. Study site location is another important factor to consider. A study sample from a hospital in China is likely to contain mostly East Asian patients, whereas a study sample from the Netherlands is likely to contain mostly Western European patients. Differences in oncologic outcomes based on ethnic background are well documented [17–19] and present a challenge to generalizability.

As the above examples illustrate, some sampling issues are common or particular to oncology. For instance, treatment at a tertiary cancer center may be different from treatment at a community center, and studies of patients from a particular geographic region may not be generalizable to the disease as a whole. Ultimately, we may not be able to fully generalize our retrospective findings to the target cancer population; nevertheless, our findings make important contributions to the understanding of that disease. The sample that one can ultimately generalize to is considered the **accessible population**.

9.4.1 Selection Bias

If the study sample is not representative of the target population and the underlying exposure-outcome relationship, then the measures of association will be biased. Selection bias exists when a characteristic of the sample makes it different from the target population in a fundamental way that cannot be ignored [20]. The selection bias can affect both who is included in the study and the likelihood of people being retained or followed up within the study. Examples include differential patient referral or diagnosis; differential screening for disease or progression; selection of a comparison group that is not representative of the target population; or differential loss to follow-up in a cohort study, such that the likelihood of being lost to follow-up is related to one's outcome or one's exposure status [21, 22].

For example, in a retrospective study, we cannot control for variations in treatment, such as which patients received treatment, when patients received it, or which surgeon operated on which patients. In other scenarios, some patients may have received an additional diagnostic test whereas others did not, or some patients may have received genetic testing while others opted out or were not even offered the test. Taking the earlier example of gallbladder cancer outcomes, if only those patients who were seen by a surgeon were included, we would have introduced a selection bias into our study if we wanted to generalize to all patients with gallbladder cancer. Patients who were seen by a medical oncologist only and were not referred to a surgeon are part of the target population as defined. Thus, if the investigator's goal is to draw conclusions about all patients with gallbladder cancer, they should obtain data from other sources, such as medical oncology, so that all patients are represented. If the data are not available, then one may want to consider restricting the sample to only surgical gallbladder patients, recognizing that this limits the generalizability of the findings to gallbladder cancer patients who underwent surgical resection.

9.4.2 Information Bias

Information bias is a major limitation of retrospective studies, as the necessary data elements were not planned in advance. For example, reported post-operative complications depend on the complication being accurately documented in the medical record, and this information may not be available in the chart. In addition, the physician may have spoken with the patient and ordered a treatment from an outside pharmacy. Also, if one is trying to determine events from hospital billing records alone, not all medical events are documented in the International Classification of Diseases (ICD)-9/ICD-10 coding system, but only those that were related to medical billing charges. Therefore, some complications may be missing or incomplete. Though information bias itself may be unavoidable, using reproducible, systematic data collection methods will decrease the impact of errors arising from retrospective data capture.

9.4.3 Recall Bias

Recall bias is a specific type of information bias pertaining to the accuracy of data recalled from a time in the past. Recall bias occurs when patients are asked to recall symptom and/or treatment details that may have occurred months or years earlier. Examples of such bias include recalling age at menarche [23, 24] or the assessment of pain after a prior procedure [25].

9.4.4 The Denominator Problem

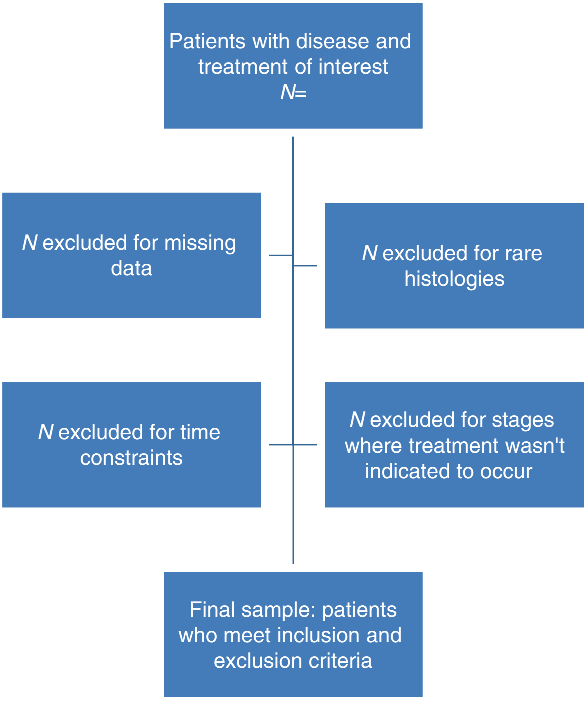
Being able to identify all patients eligible to be included in a retrospective study is a critical hurdle. Through a proxy, such as billing records, an institutional database may identify patients with a specific disease who had surgery. However, if patients were mistakenly billed for a different surgery (e.g., prostatectomy instead of prostate surgery), or the list of all possible billing codes is unknown, one could miss many patients. Further, it one's institution does not have electronic medical records or an institutional database, it may be extremely difficult or practically impossible to collect all possible patients. **Being unable to identify the number of potentially eligible patients is known as the denominator problem** [26]. This can be particularly troublesome for studies in which rates, such as post-operative complication rates or re-admission rates, need to be calculated. If not all patients were identified, these rates may be artificially higher than the true rate. The denominator problem is closely related to selection and information biases.

One common way to demonstrate how one's study sample reflects the total possible pool of patients is through flow charts. Flow charts are illustrations that demonstrate how one obtained the final sample from the initial group of patients. A flow chart enables others to get a sense of how common the inclusion criteria were and how exclusions shaped the final cohort. The following are examples of what information to include in a retrospective flow chart (Fig. 9.3):

9.4.5 Sample Selection Methods

Convenience sampling is a common selection method in retrospective research. In convenience sampling, one selects the cases that are easiest to obtain for the study. In retrospective research, this usually means that the sample is obtained from one's current institution, where one has access to the records, or is made up of patients that the researcher has treated. Because these patients are chosen for accessibility rather than representativeness, generalizability is a major problem in convenience sampling. It is important to note that, although one can employ probabilistic sampling techniques (e.g., systematic sampling) in a convenience sample to further refine it, if the larger cohort was not representative of the target population, the smaller study sample will not be generalizable either.

Fig. 9.3 Example flow chart


Patients with disease and treatment of interest *N*=
N excluded for missing data
N excluded for rare histologies
N excluded for time constraints
N excluded for stages where treatment wasn't indicated to occur
Final sample: patients who meet inclusion and exclusion criteria

The first two techniques we will describe, simple random sampling and systematic sampling, are more commonly used in epidemiologic or population-based studies, in which one has a much larger cohort than is needed to answer the research question. However, these techniques can also be applied to retrospective clinical studies in which one does have enough resources or time to collect data on all patients.

In **simple random sampling,** patients are selected from a larger sample through random selection. The number of patients and range of values to be included in the study is decided a priori, a series of random numbers is generated, and each patient is assigned a random number. In this design, each patient has an equal chance of being selected. In **systematic sampling,** the full sample is taken from a defined time period and the patients are ordered chronologically. For instance, suppose we have all colon cancer cases diagnosed in the United States from 2004 to 2014. We order them from diagnosis date starting with January 1, 2004. The study sample is then selected using a systematic periodic rule, such as each 10th patient or every other patient in the list. One may use this technique when there is a large number of cases and it would be unfeasible to collect data on all patients.

Even with retrospective studies, it is important to balance the needs for resources and time with having a sufficient number of patients to enable one to confidently answer the research question. Although both simple random sampling and systematic sampling are valid for choosing a smaller sample of patients, if one does not have enough patients with the outcome of interest in the smaller sample, the overall validity of the study findings will be questionable.

Consecutive sampling refers to selecting all patients who meet the inclusion criteria within a specific time frame. Particularly for oncology, where many diseases and treatments are rare, consecutive sampling is an extremely popular technique. However, with consecutive sampling, heterogeneity, such as differences in treatment course or in patient characteristics, is introduced, and this must be balanced against the need to have a sufficient number of patients to study. In some instances, this heterogeneity (e.g., differences in neoadjuvant treatment before surgery) can be controlled for by adjusting for these factors in the model. Another strategy for handling heterogeneity is **restriction sampling.** Restriction sampling refers to limiting the sample to individuals within a certain range of values for a confounding factor, such as age, to reduce the effect of such a factor. For instance, suppose we wanted to study outcomes for gastrointestinal stromal tumors (GISTs). Since the approval of Gleevec® (imatinib mesylate; Novartis) in 2008, neoadjuvant treatment and subsequent outcomes have changed for GIST patients. Therefore, we may want to restrict our sample to patients treated after 2008 to avoid possible confounding due to known treatment outcome differences, or we may want to separately study patients from before and after 2008. Unfortunately, restriction sampling limits the generalizability of results to those within the same range of restricted values.

9.4.6 Matching

Matching is a technique used primarily in retrospective research projects to minimize differences between comparison groups. Although one can account for the differences by including these factors in a multivariable model, in some instances matching may be more efficient. For example, if the outcome of interest is relatively rare or the target sample is small, one may not be able to incorporate all the factors in the same multivariable model. Also, the control group one can pull from may be many times larger than the target group, and data collection may be unfeasible in such a large group. By matching, one can select a comparison group that is similar enough to the target group such that the relationship between the outcome and the exposure is not attributable to the confounding factors one bases the matches on.

In **frequency matching**, one matches based on the distribution of values. For instance, if 20% of the cases were stage 1 and 40% were stage 2, one would match the control group, so that approximately 20% of the controls were in stage 1 and 40% of the controls were in stage 2. In **individual matching**, one pairs each particular patient in the target sample with a patient in the control sample. For example, if a case was a 25-year-old female patient with adenocarcinoma, the control should also be a 25-year-old female patient with adenocarcinoma. Expanding on this strategy further, either the match can be exact, where the continuous variables are identical, or one can use **caliper matching**. In caliper matching, the values are allowed to differ within a specific range, called a caliper. It is common to set the caliper to 0.25 standard deviations [27], but other calipers have been used [28]. In individually matched samples, only those patients who matched would be included in the study; all others are dropped. As a result, exact matching may lead to excessive dropping. Therefore, caliper matching is a helpful strategy, particularly with regard to covariates such as age, to prevent excessive sample loss and increase the likelihood of matching. Importantly, in individual matching, the comparison groups are no longer considered independent samples because the characteristics for the control group are dependent on what the characteristics were for the target group. Therefore, appropriate analytic methods to handle the dependency should be applied.

Propensity score matching is another technique that is used to reduce bias due to confounding variables. The propensity score is the probability of a patient receiving the treatment or experiencing the event conditional on specific factors or observed characteristics [5, 29–31]. In other words, if patients who are younger and have lower-stage cancer are more likely to receive treatment and are also less likely to die, these factors are confounding the relationship between the risk factor of interest and the outcome. Thus, these factors would be the ones to include in the propensity score. One can incorporate the propensity score into the study using several techniques: inverse probability weighting, stratification, covariate adjustment, and matching [32]. In propensity score matching, rather than matching being based on individual factors, one matches on the probability of being part of the target sample, which is determined before the matching process.

One can choose between matching with or without replacement. In **matching without replacement**, a patient can be matched to another patient only once, whereas in **matching with replacement**, a patient may be included for multiple target patients. Just as with any repeated measure, the fact that the patient appears multiple times needs to be accounted for in the analysis. Also within propensity score matching, one can choose between so-called **greedy matching** and **optimal matching**. Optimal propensity score matching chooses the match that minimizes the within-pair difference of the propensity score. In contrast, in greedy matching, a patient is first selected at random. Next, the control patient with the closest propensity score to this random subject is selected for matching. The term ‘greedy’ is used because the matching is not redone if that control subject would serve as a better match for the next randomly selected patient. That is, the patient stays matched regardless of the optimal benefit to the sample as a whole [32].

Similar to individual matching, with propensity score matching, one can set a threshold, or caliper, to decide how close the match should be. In **nearest-neighbor matching**, no restriction is made on the distance between the propensity score of the target and the control. In nearest-neighbor matching within a specified caliper distance, the propensity score is restricted by the caliper, or the maximum acceptable distance. This is similar to how calipers are used in traditional individual matching. Also, like individual matching, propensity score matching creates dependence between the cases and controls, so alternative analysis methods that account for the conditional nature of these samples should be employed [28, 29].

Propensity score matching was employed in a study on the relationship between protective lung ventilation during pulmonary resection and post-operative complications [33]. In this study, multiple factors were thought to be associated with the likelihood of receipt of protective lung ventilation and the occurrence of post-operative complications. Therefore, these factors could confound the relationship between ventilation use and complications. Matched cohorts were created from clinically relevant factors including, but not limited to, the factors that differed between patients on ventilation versus those not on ventilation. After propensity score matching, the authors assessed whether the cohorts were well balanced. The authors then performed their primary analyses using these balanced cohorts.

Control patients can be matched to target patients at a rate of one control per case, or there can be multiple controls for one case. The former is referred to as **1:1 matching**, and the latter as **1:n matching**. The overall sample size increases with additional controls, which can increase the strength or power of the findings. However, the benefit of using additional matches depends on the distribution and size of the pool of possible controls, and little added benefit may exist beyond 1:1 or, at most, 1:3 matching [24, 27, 28]. Further, increasing the number of required matches per patient increases the chance that the case may not be matched given a fixed pool of controls. It should be noted that, in propensity score matching, it is possible to have a variable number of matches for each control, which has been shown to reduce bias [34].

Once patients are matched, it is critical to check that the characteristics one matched on are balanced between the two groups to ensure the matching was done correctly. Although matching can reduce confounding between groups, it introduces an additional layer of complexity into the analysis methods. Further, matching can also account only for known, measurable confounding factors. If the groups differ in fundamental ways that cannot be controlled for, a selection bias may be present that limits the validity of one’s study.

9.5 Missing Data

Available relevant data may be limited in retrospective studies as the data were recorded or collected for clinical or other purposes outside the scope of the current study. Take, for example, a study investigating patients undergoing re-resection for incidental gallbladder cancer, which occurs when the cancer is diagnosed on pathology after a routine laparoscopic cholecystectomy. The goal of the project is to predict residual disease on re-resection using variables discovered at the earlier surgery. However, the pathology reporting and tissue collection are different for what is thought to be a standard cholecystectomy than for a known gallbladder cancer resection. For instance, the surgeons will perform a portal lymphadenectomy if gallbladder cancer is a known diagnosis. Thus, lymph node status is one factor that may be known only for those patients with cancer that is diagnosed prior to surgery. For patients in whom the lymph nodes were not removed at the incidental procedure, one cannot assume that they were negative for cancer. Additionally, patients referred for a gallbladder cancer surgery may be coming from multiple outside institutions to a tertiary cancer center or a specialist in a different hospital. Because pathology reports are not standardized across institutions or even within institutions, specific information regarding lymphovascular invasion (LVI) or perineural invasion may be missing as well.

Analyzing only those patients who have all their information is known as **complete case analysis**. Complete case analysis is a common strategy for handling missing data [35], but should only be used when the data are **missing at random (MAR)**. The term MAR refers to the situation where missing data are unrelated to the outcome. When summarizing variables, researchers should check for the proportion of cases with missing values. Unfortunately, there is no standard cut-off for the number or proportion of patients for which one should formally check how missing data affects the relationship between risk factors and outcomes. Regardless of amount, efforts should always be made to capture all missing data, which may require re-reviews of the medical records by an additional independent researcher.

Using the above example, if a small number of patients, such as one or two patients, are missing tumor stage or grade, one cannot logistically perform any formal checks, as this is too small a sample from which to make statistical inferences. Therefore, in this example, researchers should assess whether there were particular reasons for the lack of reporting. If one can reasonably assume that the missingness is a function of the retrospective nature of the study and not the result of any factors related to the study itself, then investigators can exclude these patients.

However, if patients with complete information differ from those with missing information with respect to the outcome, we cannot simply perform a complete case analysis. In the above example, suppose LVI data are missing in 25 cases, or 10% of the total study sample. In this situation, one should check whether patients with complete LVI information differ from those with incomplete LVI information with respect to the outcome, residual disease. Next, one should check whether the patients with unknown LVI status differ from those who are positive for LVI or those who are negative for LVI with respect to residual disease. If patients with incomplete LVI data differ from those with complete data, or if patients with incomplete LVI data differ from those with positive or negative LVI, then the data are not missing at random, as an underlying difference exists in those unknown cases [35, 36]. If the data are not missing at random for a particular factor, we cannot include that factor in the analysis, as our sample is not representative.

Alternatively, if the data were to be missing at random with no discernible clinical reason or observed differences with respect to the outcome, single and multiple imputation are two strategies for probabilistically assigning values to patients with missing data. Single-value imputation provides a single value, such as the mean estimate in patients with complete data, for all patients with missing data. In multiple imputations, missing values are determined based on the distribution of other known values in the data set or known values for that patient. Both of these strategies require assumptions and complex probabilistic methods, so researchers should proceed with caution when employing them [36].

Ultimately, when missing data is related to the outcome in a retrospective study, the safest strategy is to not include the factor with missing data in the model or assessment of outcome, and only include those factors where complete data is available. Although this limits the applicability of one's study to specific factors, it prevents biased estimates or erroneous conclusions. This will strengthen the generalizability to other samples and the overall validity of the study.

9.6 Considerations for Particular Oncology Outcomes

9.6.1 Peri-operative and Post-operative Outcomes

Reporting peri-operative or post-operative outcomes can be a retrospective study in itself, or it can be part of a larger study on outcomes. Peri-operative outcomes may be used in a variety of ways: for learning-curve studies to assess improvements in a new surgical technique, such as laparoscopic cholecystectomy; to assess how one surgical technique compares with another; or to see how peri-operative and post-operative diagnoses later influence survival. Peri-operative outcomes should be clearly defined prior to data collection. For instance, if an operation contains multiple procedures, the researcher needs to decide whether to consider the full operation time or only the time spent on the particular procedure. Analyzing complications has also become important to enable the generation of quality improvement programs and because, in many diseases, complications are associated with oncologic outcomes. A reasonable time period should be defined for which post-operative complications can be attributed to the surgery under study. Overall, when examining these short-term outcomes, clear definitions and methodology are essential for data accuracy and reproducibility.

9.6.2 Survival Outcomes

Survival endpoints are a critical component of many retrospective research studies. Simply estimating overall survival and other survival endpoints for specific cancers is fundamental for understanding their disease course. From these endpoints, we can establish a baseline from which to compare treatment outcomes or identify prognostic biomarkers. When the study objectives are to compare survival between two groups, it is important to report the survival data of the full cohort, as this is one way to check for sampling bias. That is, if the survival estimate of the cohort differs from previously published or clinically understood estimates, the sample may not be representative of the target population. Alternatively, there may be a problem with the way data were collected or the way time was measured.

Essential to correctly estimating survival is knowing when to start counting towards survival. This time point will depend on the patient groups one is comparing and what one is trying to estimate. Suppose we are investigating survival in patients who had laparoscopic liver resection compared with survival in those who underwent open liver resection. At first, it may seem acceptable to measure from time of diagnosis. However, not all patients underwent resection directly after their diagnosis. In fact, some patients received neoadjuvant chemotherapy, so months may have passed before these patients received surgical treatment. If one were to count the months between diagnosis and surgery as attributable to the effects of surgery, this would bias the findings in favor of patients who waited longer between diagnosis and surgery [37]. Ultimately, one should start the survival clock when the comparison of interest occurred. This allows one to attribute the time between the comparison and event outcome to the comparison of interest.

What counts as an event in a survival study is another factor to consider. As mentioned, retrospective studies suffer from information bias, so the cause of death is not always known. Although in some cancers one may be able to find the cause of death by the course of disease, this is not always the case. Also, patients may receive their primary treatment, such as surgery, at a tertiary cancer center, but then receive adjuvant chemotherapy or further treatment at a local institution, or vice versa. The investigator's current institution may possess only the death certificate or notification of death, but no notes of treatment after the initial diagnosis. This omission makes attributing survival to the cancer of interest difficult. Therefore, unless cause of death can be determined for the majority of patients who died, disease-specific survival as an endpoint should be used with caution.

When investigating disease progression or recurrence outcomes, it is important to consider how to regard death. In many studies, death will be regarded as an event. However, death, particularly in less functional or highly comorbid populations, may be due to causes other than progression of the cancer. Thus, one may want to regard death as a competing event and perform a competing risks analysis. In the first case, one assumes that a death is equivalent to a progression, or that progression had occurred at the time of death. In the latter case, one assumes that the patient's disease had not progressed and that the death prevented the progression from occurring. Assumptions are made in both cases, and which option to use depends on the disease and the study goals.

Lastly, in all studies of survival outcomes, one must consider how to count the patients lost to follow-up. In survival analyses, patients are counted in the survival models up until the point they are censored. In prospective studies, this is usually at the study close or on the off-study date. However, in retrospective studies, cutoff dating may not be so straightforward. In the United States there is no way to freely check death records for individuals, and the families of patients are not legally required to tell treating hospitals of a patient's death. Therefore, one cannot assume that all patients were alive on the last day that survival data were collected. Making this assumption would artificially prolong survival estimates. Alternatively, just because a patient was treated at outside institutions after the initial treatment does not mean that he or she was lost to follow-up on the date of the initial treatment. Assuming so would artificially truncate survival. Instead, one should use the last date a patient was known to be alive, using either clinic visit records, outside reports sent in, or phone conversations recorded with the hospital staff.

9.6.3 Treatment Response

In retrospective studies, the schedule of treatment administration and subsequent follow-up is not standardized. As a result, some patients may have received additional cycles of treatment, fewer cycles of treatment, or missed treatments in a heterogeneous fashion. Similarly, some patients may have had scans done every 6 weeks, some at 8 weeks, and some at 12 weeks. If one is looking at the time-to-treatment response, if patients' responses were not measured at the same time, then the time to response will be artificially altered due to the underlying differences in when measurement occurred. Further, treatment scheduling or drug dosing may have changed over time. To counteract this effect, one can use restrictive sampling to include only those patients with relatively homogeneous treatment schedules and response measurement samples. However, as discussed earlier, restrictive sampling limits the study's generalizability to all patients and to the real-world setting. Thus, time-to-treatment response is a difficult endpoint for a retrospective analysis. Alternatively, one could use response rate by a specific cutoff point, such as 12 weeks, and include all 6-, 8-, and 12-week assessments. Ultimately, treatment response studies must strike a delicate balance between real-world treatment experience and validity.

In the majority of prospective studies, the RECIST 1.1 (Response Evaluation Criteria in Solid Tumors) system is used to measure tumor response, which makes the findings reproducible and internally valid. In contrast, in most retrospective studies, tumor response is determined by the actual radiology report. Reporting may not be standardized and may differ across time or among different radiologists. Therefore, it may be challenging to determine what constitutes a response in a particular patient. One option for correcting this inconsistency is to have a radiologist perform a research re-read using standardized methodology. However, this option may be costly or not feasible in some institutions. When no re-read is conducted, one should record the language on the reports that constitutes a response, stable disease, and progression. These language categories should be reported in the methodology of the manuscript so that data collection is reproducible. In either scenario, deciding on a definition of treatment response before analysis begins is critical.

9.6.4 Residual Disease

In retrospective studies, residual disease status is typically obtained from the surgeon's operative report. Therefore, accuracy of this outcome is largely dependent on the consistency of definitions between surgeons. As an example, take primary debulking surgery for ovarian cancer. One surgeon may say "no residual disease present," another may say "no residual disease present greater than 5 mm," and another may say "no residual disease present greater than 1 cm." Fortunately, in primary

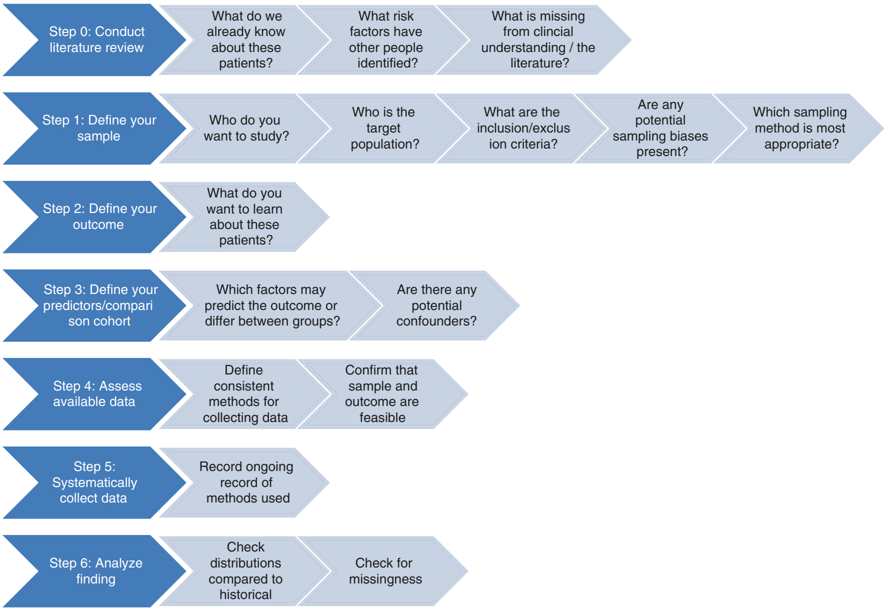

Step 0: Conduct literature review
What do we already know about these patients?
What risk factors have other people identified?
What is missing from clincial understanding / the literature?
Step 1: Define your sample
Who do you want to study?
Who is the target population?
What are the inclusion/exclusion criteria?
Are any potential sampling biases present?
Which sampling method is most appropriate?
Step 2: Define your outcome
What do you want to learn about these patients?
Step 3: Define your predictors/comparison cohort
Which factors may predict the outcome or differ between groups?
Are there any potential confounders?
Step 4: Assess available data
Define consistent methods for collecting data
Confirm that sample and outcome are feasible
Step 5: Systematically collect data
Record ongoing record of methods used
Step 6: Analyze finding
Check distributions compared to historical
Check for missingness

Fig. 9.4 Study design guide

debulking surgery, 1 cm is a generally agreed-upon cutoff, so one may assume that all these patients are free of residual disease. However, not all diseases have an agreed-upon cutoff. In other studies, one may define residual disease by site, such as not present, loco-regional, or distant residual. Therefore, to avoid biasing the findings, the best strategy for residual disease projects is to define residual disease and residual disease location sites before collecting and analyzing data. Additionally, one should consider collecting three elements for residual disease: presence/absence, size, and location. From this, one can use the data gathered to quantify the breadth of responses, while also allowing for appropriate categorization should there be disagreement in the literature. As in the above study outcomes, clear definitions and systematic data collection are the key tools for making a retrospective study internally valid and reproducible.

Conclusions

Retrospective studies allow researchers to study outcomes in a real-world setting at reduced costs compared with those for prospective trials. However, retrospective studies suffer from unique biases that researchers must pay careful attention to. It is critical that patients be selected and data captured methodically in order to make the findings internally valid, generalizable, and reproducible. We leave readers with a baseline checklist of questions to consider when designing a retrospective study, to enable them, as researchers, to better design and more easily execute these types of studies (Fig. 9.4).

References

1. Gay J. Clinical study design and methods terminology. 2010. http://people.vetmed.wsu.edu/jmgay/courses/glossclinstudy.htm. Accessed 1 May 2017.
2. Porter GA, Skibber JM. Outcomes research in surgical oncology. Ann Surg Oncol. 2000;7(5):367–75.
3. Funai EF, Rosenbush EJ, Lee MJ, Del Priore G. Distribution of study designs in four major US journals of obstetrics and gynecology. Gynecol Obstet Investig. 2001;51(1):8–11.
4. Scales CD Jr, Norris RD, Peterson BL, Preminger GM, Dahm P. Clinical research and statistical methods in the urology literature. J Urol. 2005;174(4, Part 1):1374–9.
5. Solomon MJ, McLeod RS. Surgery and the randomised controlled trial: past, present and future. Med J Aust. 1998;169(7):380–3.
6. Kærn J, Tropé CG, Abeler VM. A retrospective study of 370 borderline tumors of the ovary treated at the Norwegian Radium Hospital from 1970 to 1982. A review of clinicopathologic features and treatment modalities. Cancer. 1993;71(5):1810–20.
7. Sartwell PE. Retrospective studies: a review for the clinician. Ann Intern Med. 1974;81(3):381–6.
8. Van den Beuken-van Everdingen M, De Rijke J, Kessels A, Schouten H, Van Kleef M, Patijn J. Prevalence of pain in patients with cancer: a systematic review of the past 40 years. Ann Oncol. 2007;18(9):1437–49.
9. Markman M. A unique role for retrospective studies in clinical oncology. Oncology. 2014;86(5-6):350.
10. Hayden GF, Kramer MS, Horwitz RI. The case-control study. A practical review for the clinician. JAMA. 1982;247(3):326–31.
11. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57–61.
12. Sauerland S, Lefering R, Neugebauer E. Retrospective clinical studies in surgery: potentials and pitfalls. J Hand Surg. 2002;27(2):117–21.
13. The Cochrane Collaboration. Glossary. 2017. http://community-archive.cochrane.org/glossary/5#letterv. Accessed 1 May 2017.
14. Rochon PA, Gurwitz JH, Sykora K, Mamdani M, Streiner DL, Garfinkel S, Normand S-LT, Geoffrey M. Reader's guide to critical appraisal of cohort studies: 1. Role and design. BMJ. 2005;330(7496):895.
15. Cook TD, Campbell DT. The design and conduct of quasi-experiments and true experiments in field settings. In: Dunnette MD, editor. Handbook of industrial and organizational psychology, vol. 223. Amsterdam: Elsevier; 1976. p. 336.
16. Steckler A, McLeroy KR. The importance of external validity. Am J Public Health. 2008;98(1):9–10.
17. Ademuyiwa FO, Edge SB, Erwin DO, Orom H, Ambrosone CB, Underwood W. Breast cancer racial disparities: unanswered questions. Cancer Res. 2011;71(3):640–4.
18. Albain KS, Unger JM, Crowley JJ, Coltman CA, Hershman DL. Racial disparities in cancer survival among randomized clinical trials patients of the Southwest Oncology Group. J Natl Cancer Inst. 2009;101(14):984–92.
19. Du XL, Fang S, Vernon SW, El-Serag H, Shih YT, Davila J, Rasmus ML. Racial disparities and socioeconomic status in association with survival in a large population-based cohort of elderly patients with colon cancer. Cancer. 2007;110(3):660–9.
20. York RO. Conducting social work research: an experiential approach. London: Pearson College Division; 1998.
21. Aschengrau A, Seage GR. Essentials of epidemiology in public health. Burlington, MA: Jones & Bartlett Learning, LLC; 2013.
22. Weiss NS. Clinical epidemiology: the study of the outcome of illness. Oxford: Oxford University Press; 1996.
23. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87–91.

24. Damon A, Bajema CJ. Age at menarche: accuracy of recall after thirty-nine years. Hum Biol. 1974;46:381–4.
25. Lowe JT, Li X, Fasulo SM, Testa EJ, Jawa A. Patients recall worse preoperative pain after shoulder arthroplasty than originally reported: a study of recall accuracy using the American Shoulder and Elbow Surgeons score. J Shoulder Elb Surg. 2017;26(3):506–11.
26. Schatman ME, Campbell A, Loeser JD. Chronic pain management: guidelines for multidisciplinary program development. Boca Raton, FL: CRC Press; 2007.
27. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhyā. 1973;35:417–46.
28. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150–61.
29. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037–49.
30. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
31. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79(387):516–24.
32. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399–424.
33. Amar D, Zhang H, Pedoto A, Desiderio DP, Shi W, Tan KS. Protective lung ventilation and morbidity after pulmonary resection: a propensity score-matched analysis. Anesth Analg. 2017;125(1):190–9.
34. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat. 1993;2(4):405–20.
35. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91(1):4–8.
36. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66.
37. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1(11):710–9.

How to Design Phase I Trials in Oncology

10

Louise Carter, Ciara O'Brien, Emma Dean, and Natalie Cook

10.1 Introduction

The traditional goals of a phase 1 study are to assess the safety, tolerability and on/off-target effects of a novel investigational medical product (IMP), or combination, in trial participants. To achieve this, a 'safe' starting dose is first calculated, using preclinical data and published guidance, to estimate a maximum safe starting dose [1].

The principles of dose escalation are to escalate quickly in the absence of toxicity in order to minimise the number of subjects treated at sub-therapeutic doses. However, in the presence of toxicity, dose escalation should be slower. The key objective is to identify the recommended phase 2 dose (RP2D) of the IMP. Traditionally, the RP2D has been calculated based on the 'maximum tolerated dose' (MTD), a toxicity-based endpoint. However, with the increasing use of molecularly targeted agents (MTAs) and immunotherapy, the optimal biological dose (OBD), an endpoint based on pharmacokinetic (PK), pharmacodynamic (PD)

L. Carter, M.A., M.B.B.S., Ph.D., M.R.C.P. • Natalie Cook, M.B.Ch.B., M.R.C.P., Ph.D. (✉)
The Christie NHS Foundation Trust, Manchester, UK

Division of Cancer Sciences, Faculty of Biology, Medicine and Health,
University of Manchester, Manchester, UK
e-mail: Natalie.Cook@manchester.ac.uk

C. O'Brien, B.Sc., M.B.B.S., Ph.D.
The Christie NHS Foundation Trust, Manchester, UK

Emma Dean, B.MedSci.,B.M.B.S.,PhD.,F.R.C.P.
The Christie NHS Foundation Trust, Manchester, UK

Division of Cancer Sciences, Faculty of Biology, Medicine and Health,
University of Manchester, Manchester, UK

Early Clinical Development, Oncology Translational Medicine Unit, Astra Zeneca,
Melbourn, Hertfordshire, UK

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_10

and efficacy data is increasingly being considered in RP2D determination. Once an RP2D has been identified, a dose expansion phase may commence, allowing additional data collection on the tolerability and safety of the IMP, along with PD endpoints to demonstrate proof of mechanism and concept.

An early read-out of the efficacy and toxicity of the IMP, or combination, may be used as a benchmark for the accelerated development of immunotherapy and MTAs. An adaptive and/or modular phase 1 trial design with large dose expansion cohorts, often with companion mandatory tumour biopsies, allows robust assessment of proof-of-drug concept, drug mechanism and target engagement. A good example of this approach is the development of the third-generation epidermal growth factor receptor (EGFR) inhibitor osimertinib (AZD9291). In this phase 1 trial design, the IMP was rapidly evaluated in dose escalation cohorts of six patients, but as evidence of clinical activity emerged, expansion cohorts were opened which interrogated proof of concept by only including patients who were willing to undergo a baseline tumour biopsy for EGFR T790 status [2]. In this way the distinction and transition between the distinct phases of trial design was more seamless, thus allowing earlier approval of the IMP by regulatory authorities.

It is imperative throughout the conduct of a phase 1 trial for each trial subject to be monitored closely with a responsive and adaptive approach to reporting and managing evolving toxicity, which may be complex and idiosyncratic.

10.2 Phase Trial Designs

10.2.1 Traditional

Traditional or rule-based phase 1 design makes no assumption about the dose-effect curve of the IMP. As the trial proceeds, the dose escalation strategy and RP2D are pre-defined by the protocol before the trial begins, on the basis of observed toxicities of the investigational agent at the preceding dose level. The main advantages of rule-based methods are that they are easy to implement. However, they may be inefficient in establishing the dose-toxicity level. Also, the dose decision for the next cohort and definition of the RP2D rely on data from the current dose level and are ‘memoryless’, in that they do not include all available information.

10.2.1.1 3 + 3 Design

This simple design allows the dose-effect of the IMP to be assessed at each dose level in initial cohorts of three patients [3]. Based on the dose-limiting toxicity (DLT) data, the IMP dose can be escalated or de-escalated, using a pre-defined strategy outlined in the trial protocol, in an additional cohort of three patients. The principles of dose escalation remain the same, to rapidly escalate in the absence of toxicity, in order to minimise the number of subjects treated at sub-therapeutic doses. Some studies may initially expose a minimum number of patients in $n = 1$ cohorts to the IMP at low dose levels (see accelerated titration designs below) and use protocol-defined rule-based escalation, e.g., initial dose quadrupling until detectable PK exposure or dose doubling until the occurrence of grade 2 adverse events. Another example of a

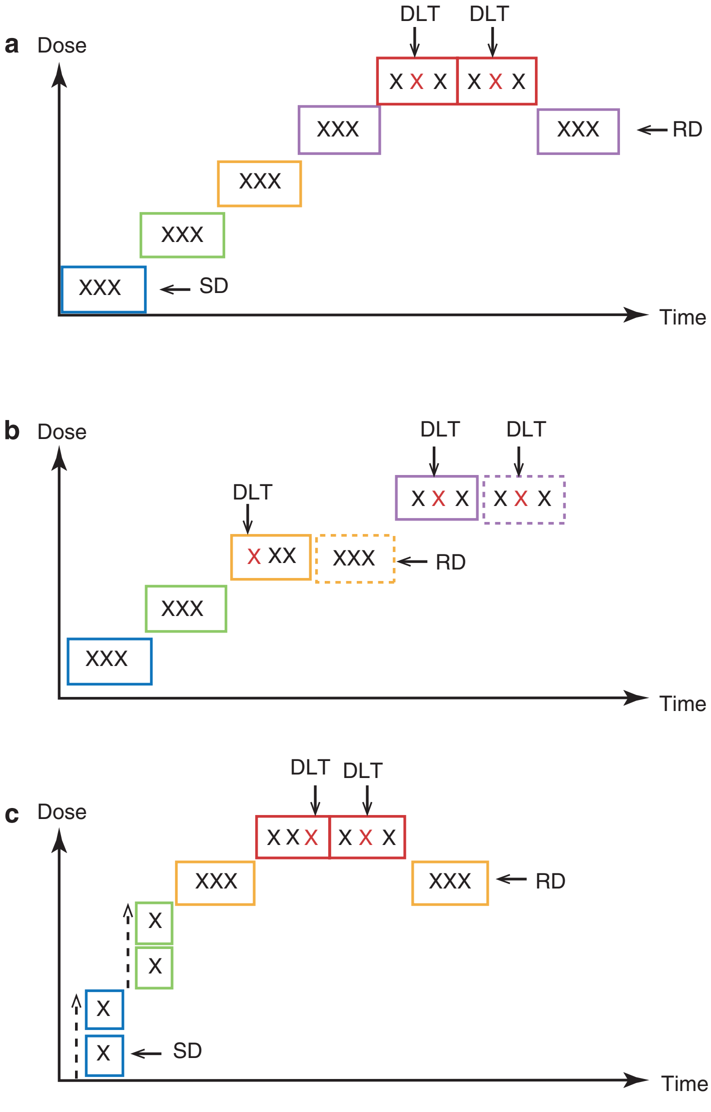


Fig. 10.1 Graphical representation of dose escalation strategies in phase 1 clinical trial design (**a**) 3 + 3 design; (**b**) rolling 6 design; (**c**) accelerated titration design. *SD* starting dose, *RD* recommended dose, *DLT* dose-limiting toxicity

pre-defined escalation strategy is the use of the modified Fibonacci sequence. This approach follows the Fibonacci sequence of fixed % dose increments, which are larger at low dose levels (and likely less toxic dose levels); dose increments are diminished as dosing proceeds and potential toxicity accumulates (typically the first dose increases by 100%, and thereafter by 67%, 50%, 40% and 30%). The RP2D is the highest dose level achieved with a specified DLT frequency (Fig. 10.1a).

The 3 + 3 design is simple for physicians to implement, transparent to use and requires minimal bio-statistical support. It is therefore one of the most widely used trial designs in Phase I studies. A disadvantage of the 3 + 3 design is that a larger proportion of trial subjects may receive a sub-therapeutic dose of the IMP than in other trial designs [4]. Furthermore, the RP2D may be relatively underestimated. The rigid dose escalation strategy in a 3 + 3 trial is a time-consuming approach.

10.2.1.2 Rolling 6 Designs

Rolling 6 designs offer the benefit of greater flexibility in dosing two to six subjects who may be recruited concurrently. The actual number enrolled is based on the numbers of patients currently enrolled and evaluable who experience a DLT and who remain at risk of developing a DLT (Fig. 10.1b).

In this way, start-stop decisions for trial accrual are more streamlined and efficient, thus improving the timeliness of trial read-out. However, similar to the 3 + 3 design, subjects in early dosing cohorts are more likely to receive a sub-therapeutic IMP dose.

10.2.1.3 Accelerated Titration

Accelerated titration is normally used for single-agent clinical trials. In this design, one patient is treated per dose level (cohort) with fixed dose escalation increments (e.g., 40% or 100%) until a DLT (grade 3 or 4 toxicity) or pre-defined trigger (PK exposure levels or grade 2 toxicities) is observed. Thereafter, the trial proceeds by utilising a traditional 3 + 3 design or an adaptive/model-based design to closely monitor the safety and tolerability of the IMP at higher dose levels (Fig. 10.1c). Intra-patient dose escalation may also be permitted.

Accelerated titration aims to reduce the number of trial participants treated with a sub-therapeutic IMP dose and, as a general rule, explores fewer dose levels. Consequently, data is collected more rapidly than in the 3 + 3 or rolling 6 designs, but at a cost of reduced capacity for late toxicity assessment at the lower dose levels. In addition, there is greater likelihood of trial participant exposure to grade 3 or 4 toxicity.

10.2.2 Adaptive Designs

An adaptive or model-based phase 1 design determines a pre-defined dose-effect curve prior to trial recruitment. This dose-effect curve is modified as the trial proceeds, based on the toxicity of the IMP. This real-time adaption of the trial design as a rule requires excellent bio-statistical support for successful implementation. More patients usually receive doses near the MTD with model-based designs, which can also provide a measure of precision and incorporate ‘memory’ or safety information from all patients regardless of dose level [5]. The adaptive design is preferable for the conduct of more complex combinations of IMPs clinical trials, or where it is preferable to assess efficacy and toxicity simultaneously to allow rapid assessment and development of an IMP. It is for this reason that many trials methodology consortia now recommend model-based designs as the preferred design. These recommendations are summarised in the following document: “A quick guide why not to use A + B designs” (MRC Hubs for Clinical Trial Methodology Research) [6].

10.2.2.1 Continuous Reassessment Model (CRM)

In the CRM design, preclinical data is used to predict the probability of dose-toxicity in a defined (fixed) sample of trial participants at different IMP dose levels to guide initial dose escalation. However, patients are generally treated at a dose thought to be close to the MTD. As the clinical trial proceeds, Bayesian probability statistics are used to model and re-model the relationship between IMP dose and cumulative toxicity and thereby modify the dose escalation strategy until a pre-specified criterion is met, at which point the trial terminates.

The CRM design relies on preclinical data, which may not always be readily available, to model dose toxicity at lower dose levels. There is the risk that the MTD may be over-estimated by this design, exposing trial participants to a higher risk of toxicity. Robust and timely statistical support is a pre-requisite for this trial design.

Modified CRM designs use a more conservative initial dosing strategy with single dose level increments per cohort, increasing the dose by only one pre-specified level at a time, and with larger participant cohorts within the trial compared with standard CRM. Stopping rules are pre-defined within the trial protocol and are not based on sample size, in contrast to standard CRM. Where a DLT occurs, the dose delivered to the next participant may not be escalated. Modified CRM enhances safety compared with CRM and may improve efficiency of trial conduct. Time-to-event (TITE) CRM modifies the standard CRM design to improve the capture of cumulative toxicity data. The observed ‘event’ refers to time to toxicity, and therefore toxicity from all patients recruited into the trial is captured and incorporated into the continuous dose modelling strategy in real time. The contribution of each trial participant to the dose-toxicity curve is weighted dependent on DLTs observed (or not observed) and the time the participant has been on the clinical trial (drug exposure). The TITE CRM is a preferable design for capturing chronic toxicity data in an IMP clinical trial.

Escalation with overdose control (EWOC) is a further modification of the CRM design [7]. The emphasis with this trial design is safety, with analysis of toxicity data from preclinical studies and analysis of clinical dosing after each patient (rather than cohort) receives a specific dose level to assess the probability of overdose at each dose escalation decision point. This design is resource intensive, as it requires on-site bio-statistical support for successful trial conduct.

10.2.2.2 Modified Toxicity Probability Interval (mTPI)

The mTPI design uses a Bayesian framework and hierarchical categorisation (underdosing, proper dosing and overdosing) to compute dose escalation, based on the interval between the toxicity rate of each dose level and target probability [8]. Simply, where the toxicity interval is within the ‘underdosing’ category, the recommendation is to escalate the dose; if toxicity falls in the ‘proper’ dosing category, the dose is maintained, whereas if toxicity falls in the ‘overdosing’ category, the dose is de-escalated. Therefore, all dose decisions for a trial can be pre-calculated and plotted on a 2 × 2 table (number of toxicities by number of participants treated per dose level) to allow the trial physician to select the best dose escalation strategy as the trial proceeds. The mTPI adaptive model requires bio-statistical support at trial set-up, but thereafter the trial may proceed based on pre-determined rules for dosing. The mTPI model minimises the risk of grade 3 or 4 toxicity and the risk of dose delivery exceeding the MTD to trial participants.

In the mixed effect proportional odds model (mixed effect POM), toxicity is graded and collated to allow calculation of the odds of toxicity per cycle. The RP2D is defined as the dose associated with a pre-defined probability of severe toxicity per cycle.

10.2.2.3 Fractional Dose-Finding Methods

Fractional dose-finding trial methodology aims to offset the bias towards acute toxicity in dose finding with the assimilation of late toxicity data [9]. Time to toxicity is calculated utilising a Kaplan Meier plot for recorded toxicities or censored observations (where a DLT has not occurred) for each trial participant. The fractional contribution of each trial participant to the dose-toxicity curve is weighted based on their time on IMP to inform dose escalation. Censored observations in trial participants where toxicity is yet to be observed are weighted towards the right to counteract bias.

10.2.3 Pharmacokinetically Guided Dose Escalation (PGDE)

The PGDE trial design makes the critical assumption that the plasma dose concentration in animals can predict the dose-toxicity relationship of the IMP in humans and thereby guide dose escalation strategy. For this design, robust preclinical data is necessary and regular PK assessment is built into the conduct of the clinical trial. Drug exposure, as defined by the area under the curve (AUC), is measured at the first dose level and thereafter dose escalation proceeds according to the distance to the target AUC. This may proceed by a factor equal to the square root of the target AUC/initial AUC and/or by a modified Fibonacci sequence.

The PGDE trial design is particularly helpful for clinical trials of MTAs where classical monotonic dose relationships are not observed. This approach allows an objective estimation of the OBD using a real-time approach. The use of preclinical data assumes that dose-toxicity relationships are conserved across different species. The PGDE design does not account for non-linear PKs and offers little option for exploring different dosing schedules within the clinical trial. Dose escalation is guided by real-time PKs and is therefore resource intensive and impractical to run on some phase 1 units. Furthermore, dose escalation is guided by acute rather than chronic toxicity data capture. PK-guided dose escalation may be adversely affected by inter-participant variability in small patient cohorts.

10.3 Choosing a Starting Dose

The selection of a safe starting dose in first-in-human (FIH) studies is a critical step and must be weighed against the ethical concerns of treating early trial participants with an IMP well below any therapeutic benefit. As a general rule, the IMP starting dose is biometrically scaled from preclinical toxicity data obtained from rodent and non-rodent (more biologically applicable to humans) species. By convention, the starting dose in FIH trials has been 1/10 (lethal dose; LD10) of the highest non-severely toxic dose (HNSTD) in rodents or 1/6 of the HNSTD in non-rodent species (Table 10.1). Where a novel IMP is considered at high risk of significant human toxicity, rather than starting with the highest dose that is considered safe, starting dose calculation also considers the lowest active dose or 'minimum anticipated biological effect level (MABEL)'. Useful guideline documents for dosing cytotoxic IMPs and small MTAs in early phase trial designs have been published by the Committee for Medicinal Products for Human Use (ChMP) 2007 and the United States Food and Drug Administration (FDA) (S9 ICH) 2010 (Table 10.1) [10, 11].

An important point to consider is that preclinical toxicity data may not adequately characterise newer cancer agents such as monoclonal antibodies, antibody drug conjugates and immunotherapies where MTD is not attained in the preclinical work-up of the IMP. In this circumstance, toxicity data should be considered alongside biological endpoints such as PK and PD to determine dosing strategy in FIH clinical trials [12, 13].

10.4 Endpoints in Early Phase Trial Design

10.4.1 Toxicity Endpoints

The traditional clinical endpoints for phase 1 trials are safety, tolerability and, specifically, dose-limiting toxicity, usually determined during the first cycle of treatment. This is generally defined as grade 3 or 4 toxicity, but may include cumulative

Table 10.1 Principles of dose finding in first-in-human clinical trials (modified from FDA guidelines)

Step 1	Determine no observable adverse effect level (NOAEL; mg/kg in toxicity studies)
Step 2	Convert NOAEL from the most appropriate species to a human equivalent dose (HED)
Step 3	Select HED, taking into account additional factors such as metabolism, receptors and binding epitopes
Step 4	Apply a safety factor (default is ≥10-fold) and divide HED by that factor = maximum recommended starting dose (MRSD)
Step 5	Adjust MRSD based on the pharmacologically active dose

NOAEL no observable adverse effect level, *HED* human equivalent dose, *MRSD* maximum recommended starting dose grade 2 toxicity lasting more than 7 days despite supportive medications. The toxicity grading is performed with the National Cancer Institute Common Toxicity Criteria for Adverse Events (NCI-CTCAE), which are continuously updated based on new data about the toxicity of new pharmacological agents. There is a direct biological correlation between increasing dose of IMP, cytotoxicity and efficacy in in-vitro and in-vivo model systems. However, finite dose escalation is governed by subject tolerance, which can be read-out objectively as graded toxicity assessment and MTD. For some IMPs where chronic toxicity is anticipated, the DLT assessment window may be extended or, for rare serious expected toxicities, multi-institutional trials may evaluate an IMP in larger patient numbers, but provide data for individual investigators who have less familiarity with the toxicities. The phase I physician should also be mindful of mechanistic toxicities that may pertain to a class of MTAs, e.g., nail and hair changes with drugs that modulate calcium and phosphate; hypertension and proteinuria with anti-angiogenics; hyperglycaemia with PIK3CA inhibitors and immune-related side-effects with checkpoint inhibitors, as well as idiosyncratic reactions. Alternative tolerability endpoints include time to onset of AEs/DLTs, or the proportion of patients requiring a dose reduction or interruption.

Importantly, it should not be assumed that MTD is interchangeable with OBD (see section below) and in fact, there may be a significant discrepancy between these two endpoints, which may significantly impact toxicity (acute and chronic) and therefore, the tolerability and safety of the IMP in a real-world setting. This is particularly relevant for MTAs, where dosing may be continuous and long term. The effects of chronic toxicities such as fatigue and rash are often underestimated in traditional assessments of toxicity, and these effects are of particular significance in choosing the correct dose for a continuously dosed MTA [14]. The multiple tyrosine kinase inhibitor cabozantinib is associated with a number of toxicities, including diarrhoea, fatigue and plantar-palmar erythrodysesthesia, often necessitating dose reductions [15]. The use of lower doses than the MTD of cabozantinib is not associated with reduced exposure, suggesting there are doses which remain biologically active but are more tolerable [16]. In reality, many non-optimal doses are taken into late development, with a high rate of dose interruptions and reductions observed in registration trials and post-marketing. Monoclonal antibodies often do not produce the traditional toxicities associated with DLT definitions and so ongoing careful consideration of dose is required.

The therapeutic window of a drug refers to the range of doses or concentrations that are both efficacious and tolerable (Fig. 10.2). This index defines the margin of safety for a drug, establishing the ratio of the dosage that produces toxicity in 50% of treated patients to the dosage that produces a desired effect in 50% of the patients. In drugs with a narrow therapeutic window the MTD and the OBD are likely to be similar. However, for drugs with a wide therapeutic window the potential benefits of determining the OBD as opposed to the MTD are more marked. Therefore, it is now good practice to assess cumulative toxicity alongside biological endpoints of IMP activity, such as target validation, PD and PK, beyond cycle one of a FIH clinical trial. This information can inform dose escalation/de-escalation decisions of later trial cohorts. Without the characterisation of MTD and

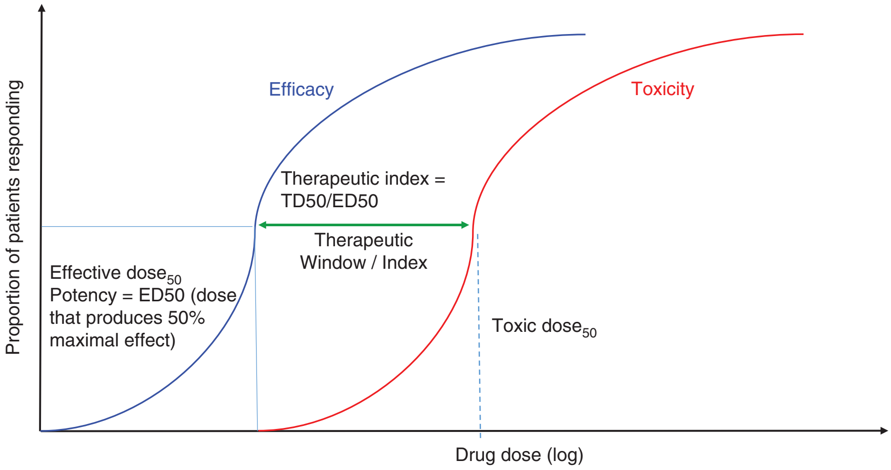


Fig. 10.2 The therapeutic window of a drug. The therapeutic window of a drug refers to the dose range for a drug between the doses that are effective and the doses that are toxic. The therapeutic index refers to the ratio of the dose of the drug that produces a specific rate of toxicity, e.g., toxic dose in 50% of the subjects, to the dose of the drug that produces a specific rate of a desired pharmacological change, e.g., efficacious dose in 50% of subjects. *TD50* toxic dose in 50% of subjects, *ED50* effective dose in 50% of subjects

OBD, further investigation of the mechanism of action of the drug is mandatory prior to further IMP development.

In addition, integration of patient-reported outcomes into clinical trial design can be performed to increase the speed and accuracy of data collection [17]. Similarly, real-time data collection of adverse events may be uploaded into clinical software platforms to improve the quality of data capture and the agility of adaptive trial design [18]. This enables data operators to detect trends, using software to flag and alert investigators, in real-time, of anomalies and also enables the visualization of data using heat-maps.

10.4.2 Biological Endpoints

Chemotherapy, through targeting rapidly dividing cells, has a monotonic relationship between dose and response, supporting the rationale of using the MTD to determine the RP2D. MTAs, in contrast, preferentially target an aberrant pathway within the cancer cells, with relative sparing of the normal tissues, such that toxicity and efficacy are not intrinsically linked to dose. It has therefore been proposed that the OBD may be a more appropriate endpoint than MTD in phase I trials of MTAs. The OBD is the dose of a drug associated with the most favourable change in a selected PD, PK or functional imaging biomarker. However, for the OBD to be a suitable endpoint, a number of criteria need to be met; namely, there needs to be access to tumour tissue or a surrogate tissue for analysis from patients in the trial.

The assay should be a reliable assay available to measure the desired effect and the optimal extent to which the target must be inhibited by the drug, or the optimal PK threshold must have already been established from prior studies.

The OBD of an individual drug will be determined by its mechanism of action, but will also be influenced by other factors such as tumour type, toxicities and the population in which it is trialled. Multiple factors influence the systemic exposure to drugs, producing inter and intra-patient variability, such as pharmacogenetics and physiological factors including age, sex and race. Although some of these can be controlled in trials, such as restricting the use of concomitant medications and complementary therapies, the variability in systemic exposure will influence the ability of trials to determine the OBD of a drug. The population of patients in which a drug is trialled will also influence the determination of the OBD for MTAs, with trials of phase I drugs in a molecularly selected population of patients potentially able to assess the OBD more accurately through assessment of its mechanism of action than in trials in unselected populations. For example, trials of the EGFR tyrosine kinase inhibitor erlotinib in EGFR mutant non-small cell lung cancer (NSCLC) have shown that an impressive tumour response could be seen at doses between 25 and 75 mg [19, 20]. However, the initial phase I trials in unselected populations led to 150 mg being the RP2D. The optimal dose may also be tumour-specific, rather than the one-size-fits-all approach of determining chemotherapy RP2D. For example, the recommended dose of bevacizumab, an anti-vascular endothelial growth factor (VEGF) monoclonal antibody, differs in renal cell carcinoma, colorectal cancer, breast cancer and NSCLC [21–24]. Although tolerability assessments are not used in determining the OBD of a drug, they are still relevant in determining the RP2D.

Although it is an enticing concept, using the OBD as an endpoint for phase I trials presents a number of challenges. First, a strong scientific rationale is a prerequisite when exposing patients to biopsies. Sufficient access to tumour tissue or surrogate tissue is required to enable an accurate relationship between dose and response to be established. Paired tumour biopsies, pre- and post treatment, represent the gold standard for analysing tumour biomarkers [25]. However, tumour biopsies are invasive and associated with risks to the patients; so, particularly at low drug doses, there are questions about the ethics of their use. Surrogate tissues such as peripheral blood mononuclear cells and hair follicles are used, but they may not accurately reflect changes in the rapidly dividing tumour, particularly when targeting a pathway that is aberrant in the tumour but not within the surrogate tissue. Robust preclinical data must be available to support the mechanism of action of the drug and the changes produced within the tumour. It must also have been established how much target inhibition is required to produce the optimal effect on the tumour. The assays used to assess changes in the tumour to determine the OBD, such as target inhibition, have to be reliable and reproducible so they are fit for purpose to be used as the basis of dose decision-making. Given the heterogeneity of tumours, how much tumour inhibition within a single biopsy can be extrapolated to represent the entire tumour must also be considered [26]. Unfortunately, both PD and PK biomarkers are often not sufficiently robustly tested in preclinical tumour models and the relationship between them and the alterations in tumour growth rates are not well characterized, impacting the ability to accurately determine OBD [27].

A number of trial designs utilising OBD to determine the RP2D have been proposed [28–30]. But the difficulties associated with implementing OBD as the primary endpoint in clinical trials have resulted in it being used less commonly in phase I trials as opposed to toxicity endpoints. An analysis of the use of biomarkers in phase I trials from 1991 to 2002 concluded that, although biomarkers were used to support dose selection in 13% of the trials in which they were analysed, in only one case was an OBD used as the primary determinant of RP2D [31]. Subsequent analyses have suggested that the role of the MTD in defining the RP2D may be declining, particularly in trials of MTAs, with just 58% of the trials of MTAs assessed using the MTD as their primary endpoint [32]. Also, increasing importance is being placed on correlative biomarker studies within phase I trials, with an emphasis on tissue collection. There was an increase, from 14% to 26%, of phase I trials that included biomarkers from 1991 to 2002 [31]. The PK-PD relationship is now frequently investigated in phase I trials, which is likely to lead to increased knowledge about the biology of a specific dose rather than just its toxicity. With the increased drive for efficient drug development and the use of combinations of drugs to optimise response and overcome resistance, accurately establishing the OBD of drugs in phase I trials is likely to be of increased importance.

10.4.3 Efficacy

Increasingly, phase I trials incorporate efficacy endpoints, such as the response rate based on radiological assessment and/or the PD response to set go-/no-go criteria to enable decisions to be made on whether to expand the trials or terminate their development. In fact, regulatory authorities have recently recognised efficacy data from phase I expansion cohorts in licensing applications, e.g., for osimertinib [2, 33]. However, given that the majority of phase I trial participants have progressing disease at study entry, response data may underestimate clinical activity in this setting. Other criteria which have been suggested to assess efficacy include the clinical benefit rate (CBR; which includes stable disease) or comparing the rate of progression on the trial with the rate of progression pre-study by using two interval baseline imaging assessments, i.e., the ratio of progression-free survival (PFS) on study compared with PFS pre-study (PFS2:PFS1) [34].

10.5 Trial Design in the Era of Personalised Medicine

The drive towards personalised medicine is influencing all aspects of drug development, including phase I trials. The limitations of using the histology of a tumour alone to determine appropriate treatment options have become increasingly clear. The molecular basis of cancers is continually being revealed, which, combined with

Table 10.2 Characteristics of and differences between basket trials, umbrella trials and expansion cohorts

Trial characteristic	Basket trial	Umbrella trial	Expansion cohorts
Tumour types/histologies	Multiple	Single	Selected
Molecular aberrations	Single	Multiple	Selected
Drugs tested	One	Multiple	One
Randomised or non-randomised	Either	Either	Non-randomised
Outcomes	RR or PFS	RR or PFS	Efficacy and toxicity
Examples	MATCH [44] SHIVA [74]	BATTLE [75] MATRIX [76]	Keynote-001 [39] Checkmate-003 [77]

RR response rate, *PFS* progression-free survival

the increased availability of next-generation sequencing (NGS), has led to a focus on personalised medicine. As the mechanism of actions of MTAs is the targeting of aberrant pathways in tumours, there is increased recognition that non-selective recruitment of patients to phase I trials may not be the optimal method for their testing. When there is robust preclinical data about the mechanism of action of a drug and the accuracy of a predictive biomarker, patient enrichment in both the escalation and expansion phases of a trial should be considered. This is particularly of significance when an OBD is the primary endpoint, as opposed to classical toxicity-based endpoints. With phase I trials increasingly being used to seek efficacy signals, particularly in the case of rare genomic aberrations, patient selection for trials of MTAs is important. These factors have all driven the heightened interest in genomic-based clinical trials (Table 10.2).

10.5.1 Expansion Cohorts

Phase I clinical trials have developed an increasing focus on enrichment for specific patient populations, either those with specific tumour histology or those with tumours with specific molecular aberrations, as opposed the traditional all-comer trials. This is leading to significant changes in how phase I trials are designed and how phase I units are run. Large multi-arm phase 1b expansion cohorts are growing increasingly common. These expansion cohorts are commonly performed to gain further insight into the efficacy and safety of a drug at the RP2D, optimising the volume and quality of data available. Manji et al. investigated the use of expansion cohorts in single-agent phase I cancer trials, identifying 149 trials which included expansion cohorts from the 611 trials published between 2006 and 2011 [35]. They found a significant increase in the use of expansion cohorts in single-agent phase I trials, from 12% in 2006 to 38% in 2011, particularly in multi-institutional trials and those with MTAs. Objectives for the expansion cohort, including safety, efficacy, PK, PD, and patient enrichment, were reported in 74% of these trials. The RP2D was modified in 13% of the trials and new toxicity described in 54% of the trials in which safety was a stated objective.

Expansion cohorts are also being increasingly used to streamline drug development, reducing the gaps between trials and aiding with "go/no go decisions" prior to embarking on phase II/III trials. The FDA created a "breakthrough therapy" designation, which can lead to accelerated drug approval pathways [36]. This has resulted in increased focus on efficacy as an endpoint in phase I trials, thus driving larger expansion cohorts [37, 38]. A noteworthy example is pembrolizumab, an anti-PD-1 monoclonal antibody, which received accelerated approval from the FDA in 2014, 3 years after clinical development began, on the basis of efficacy data from a phase Ib expansion cohort of 135 patients [39]. From their study of expansion cohorts, Manji and colleagues did, however, note that antitumour activity was unlikely to be seen in the expansion cohort if it was not present in the escalation cohort [35]. The enrichment of specific patient subgroups in the dose expansion cohorts may therefore provide confirmatory data about those patient populations who are sensitive or resistant to the drug under investigation in addition to data generated in the escalation cohorts. In future, large dose expansion cohorts are likely to be an increasingly common feature of phase I trials. Protocols will need appropriate statistical input to ensure that they are powered sufficiently to meet their objectives but remain flexible so they can adapt to emerging data from the dose escalation of these novel agents. Phase I trial units will also have to adapt to this different style of trial, which will require increased rapid access to specific patient subgroups and large multicentre collaborations.

10.5.2 Genomic-Based Trials

Genomic-based trials use molecular abnormalities identified within tumours to select or stratify patients for clinical trials. Different trial designs have been explored for genomic-based trials, including basket and umbrella studies. Some trials include a standard-of-care arm to act as a comparator for the molecularly selected arms. Bayesian adaptive trial designs have been explored to allow a trial to adapt to emerging data [40, 41]. In the ISPY 2 trial (The Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Biomarker Analysis 2) (NCT01042379), patients with primary breast cancer are adaptively, randomly assigned to a screening process to evaluate novel agents and their associated biomarkers in combination with standard neoadjuvant chemotherapy. Experimental arms will be added or removed from the trial depending on emerging data, with the rate of pathological complete response as the primary endpoint.

It is critical in genomic-based studies that the predictive biomarkers used to select patients for treatment are "fit for purpose" to avoid misclassifying patients and undermining the purpose of the trial. When genomic-based trials only recruit biomarker-positive patients they generate no new data about the accuracy of the predictive biomarker, unlike those trials designed to include a biomarker-negative cohort. To maximise the potential of genomic-based phase I clinical trials, combinations, rather than just monotherapy arms, need to be investigated, though acknowledging this significantly increases the complexity of these trial designs.

Genomic-based studies have many potential advantages as a design for clinical trials. They allow the relationship between molecular aberrations, histology, other clinical features and response to MTAs to be explored. They have the potential to provide information about both the biology of response and resistance to novel drugs. They allow patients with rare tumours to access clinical trials and also allow drugs targeted at rare molecular aberrations to be investigated. These trials are, however, not without their difficulties. As with all studies of MTAs, strong preclinical and potentially clinical evidence that the molecular aberration being targeted is acting as a driver is required, though it is often challenging to obtain. As many molecular aberrations of interest are rare, these trials have to be performed collaboratively, often at multiple sites, to recruit sufficient patient numbers. This limits the knowledge of any one investigator about the drug under investigation, unlike more traditional phase I trials run at small numbers of sites. Genomic-based studies also rely on the infrastructure for the molecular analysis to be set up, and as the results of this analysis determine entry to these studies the turnaround time must be short, necessitating efficient processes. Given the rarity of some of the molecular aberrations under investigation, large screening programmes to identify relevant patients may be required. National or international registries of patients with specific molecular aberrations may enable patients to be rapidly identified for these types of trials in the future. Project Genomics Evidence Neoplasia Information Exchange (GENIE) is an international data-sharing project that will collate clinical-grade cancer genomics and clinical outcomes from eight phase I centres to provide data to drive forward clinical and translational research [42]. Project Genie highlights the collaboration required between multiple sites to generate the infrastructure and databases to support complex genomic trials, as well as highlighting the potential strength of the research such collaborations could produce.

10.5.2.1 Basket Studies

Basket trials comprise a clinical trial design in which one or more drugs are tested on patients with specific molecular aberrations of interest, irrespective of the histological subtypes of the tumours. Vemurafenib was tested in non-melanoma V600E mutant patients in a phase I basket study, with 42% response rates in the NSCLC cohort and 43% response rates seen in the Erdheim–Chester disease/Langerhans'-cell histiocytosis cohort [43]. An ongoing example of basket studies includes the NCI molecular analysis for treatment choice (MATCH) study (NCT02465060), in which patients are screened for actionable mutations and allocated to one of 24 arms in a histology agnostic process [44]. Basket trials, as they recruit patients with multiple different histologies, often feature stratification by histological subtype in their design. In many basket trials, in addition to cohorts for the commoner tumour types present, a cohort for patients with all other tumours with the molecular abnormality is included and so basket trials represent an opportunity for patients with rare tumours to access clinical trials. Stratifying by histology reduces the possibility of a false-negative result in which a benefit in one group of patients is missed when viewing the entire trial population as a whole. Differences in outcomes by tumour types may reflect distinctions in biology, where a given mutation may be a driver in one tumour but not in another. Cohorts with initial signals of efficacy can be expanded to enrol more of the patients with the histology of interest, or those cohorts with poor results can be closed, allowing the trial to adapt to emerging data. Basket trials can be used for hypothesis-testing prior to larger-scale studies. However, in some cases, basket studies can be used to gain regulatory approval, as for example, a phase II open-label, global, multicentre basket trial of entrectinib in patients with NTRK/ROS1/ALK fusion that is being carried out with registration intent (NCT02568267).

10.5.2.2 Umbrella Studies

In contrast to basket trials, in umbrella trials patients with one specific histological subtype are recruited and matched by an algorithm to different drugs based on the molecular aberrations their tumour contains. Umbrella trials allow multiple MTAs to be tested for a specific histological indication, rapidly highlighting those with and without potential efficacy to be prioritised for further trials. An example of an umbrella trial is the Lung MAP trial (NCT02154490), in which patients with squamous NSCLC are matched to a number of therapies based on the molecular aberrations they contain [45].

10.6 Combination Therapy in Early Phase Trials

Cancers are recognised to be increasingly complex systems with multiple aberrant pathways and processes driving their growth. Both intrinsic resistance to therapies; for example, due to redundant pathways, and acquired resistance, which develops under the pressure of treatment, cause significant challenges for drug development. Targeting just one aberrant pathway through the use of a single-agent MTA has therefore been shown to often result in only modest benefit in the majority of patients. The focus has increasingly switched to combining different therapeutics to maximise their responses. As the majority of chemotherapeutic regimes that are curative are combinations of different chemotherapies, it is not surprising that combinations of other modalities of therapies are also required [46].

Combinations of agents must be chosen with the rationale that they will be superior to either single agent alone. Additivity refers to the scenario in which each drug tested has clinical activity alone, and when combined, the activity is equal to the sum of activity of both single-agent drugs. In contrast, synergy refers to the clinical activity of the combination being greater than the sum of the activity of each single-agent drug. Synthetic lethality refers to the simultaneous loss of function of two genes resulting in cell death, whilst the loss of only one of the genes results in a viable cell. In drug development the classic example of this is the susceptibility of BRCA1/BRCA2-deficient cells to treatment with a poly (ADP-ribose) polymerase (PARP) inhibitor [47]. Antagonism refers to the scenario in which the combination therapy has less activity than the sum of the activity of each drug alone.

Given the massive number of oncology drugs in testing, the number of potential combinations of MTAs, chemotherapies, immunotherapies and radiotherapy is overwhelming. Through knowledge of the biology of cancer, computational network-based algorithms can examine gene regulation, signalling pathways and the interplay between the tumour microenvironment and the immune system to identify new therapeutic targets or resistance mechanisms [48, 49]. High-throughput system-based approaches can be used to identify novel, and in some cases, unexpected combinations of drugs for testing [50–53]. The American NCI assessed combination trial designs, and central to their recommendations was the need to have a strong scientific rationale for the hypothesis being tested as the basis of the design of the trial [54].

Having identified a combination of drugs to trial, determining the optimal dose of the combination of drugs has complexities additional to those in dose-finding studies for single agents. The ultimate aim of a combination dose-finding trial is to determine the most active combined dose level that remains tolerable. A number of rule-based strategies have been proposed for dose finding for combinations of treatments. Each agent could be dose-escalated in alternate cohorts to reach the maximum possible dose level of both drugs. Alternatively, both drugs could be simultaneously escalated in a number of cohorts to determine the maximum combined doses possible. One of the two drugs could be fixed at the MTD as a single agent whilst the second drug is gradually dose-escalated to the maximum tolerated level alongside. Finally, one drug could be escalated through a number of dose levels towards the MTD whilst the other drug is kept at a low level. These last two methods require a prioritisation of the dose of one drug, felt to be most critical to be at MTD or the most active. If it is suspected there is a limited risk of overlapping toxicity or drug-drug interaction from the two agents being tested, only a limited number of doses close to the single-agent MTD of both doses may need to be tested.

Bayesian model-based designs also are used to optimise dose escalation studies of combination therapies [55–57]. These designs are independent of assumptions about the best dose combinations. Some methods incorporate all available toxicity data to determine the appropriate doses to trial in the next cohort, which could be either an escalation or de-escalation of dose [58]. Continual reassessment models are able to evaluate all available toxicity data, including late or chronic toxicity, which is of particular relevance for trials of MTAs [59]. Models can also incorporate efficacy data in addition to toxicity data to determine a combination that is both safe and efficacious. PK data for potential drug-drug interactions can be analysed, particularly if single-agent run-in periods are included in trial designs when making dose decisions. Despite the potential benefits of using a model-based system, one analysis of combination trials determined they were used in only 4% of trials, in part perhaps because of the need for continual biostatistical input [60].

The design of combination studies becomes even more complex when scheduling is considered. Adjuvant tamoxifen delivered simultaneously with anthracycline-based adjuvant chemotherapy was found to be inferior to this treatment delivered sequentially, highlighting the importance of scheduling [61]. Continuous dosing may not be the optimal schedule for many drugs, particularly those with a narrow therapeutic window [62]. To achieve better tolerability and efficacy of combination treatment, multiple schedules may need to be assessed, such as pulsed schedules, e.g., 7 days on and 7 days off, or drug holidays may be required, based on the preclinical data available with regard to the drug exposure required for an antitumour effect. As chronic toxicities are of particular concern for MTAs, pulsed schedules may decrease toxicity whilst maximising the antitumour effect. Scheduling is also of significance when considering the combination of drugs required to overcome resistance mechanisms; namely, whether the drugs should be given in combination from the start of treatment or whether the drug combination should be given at the time of emergence of a resistant clone.

Significant challenges remain for the design and implementation of combination studies. Robust preclinical data needs to be available to support the combination being evaluated, and this is often lacking with models struggling to identify the synergistic or additive benefits of combinations. Many of the chemotherapy combinations and chemotherapy/MTA combinations tested to date were empirically chosen with the key criteria of tolerability and lack of PK interactions, rather than the choice being driven by a scientific hypothesis [63]. Preclinical models are often unable to accurately predict overlapping or supra-additive toxicity, which represents a significant stumbling block for the development of many combinations of therapies [64]. Despite the significant preclinical rationale for the combination of MEK and PI3K pathway inhibitors, clinically this combination has proven to have significant dose-limiting toxicities, such as rash and diarrhoea, limiting its potential [65]. The lack of appropriate predictive preclinical models is particularly acute in the investigation of immune oncology agents, given the species-specific differences in immune response [66]. Immune-related toxicities are also often not dose dependent and a number of agents in single-agent studies did not reach the MTD, adding to the challenge of identifying appropriate doses for combination studies. Pharmacokinetic interactions also need to be considered when designing trials to investigate combination regimens. Using single-agent PK data, investigators concluded that there would be no drug-drug interaction for a combination regimen of pazopanib and lapatanib [67]. However, a subsequent PK analysis at the RP2D revealed an interaction resulting in a suboptimal dose of lapatinib, highlighting the importance of accurate PK data [68]. Despite the challenges implicit in combination trials, these trials represent an essential area of investigation for oncology, given the raised response rates and the tantalising possibility of curing advanced disease [69]. Focusing on how to overcome the additional hurdles of early phase combination trials is therefore of increasing importance.

10.7 Future Directions for Early Phase Trials

Preclinical data will continue to be the bedrock on which clinical trials are designed, but the data clearly needs to be improved to meet the evolving needs of trial design. Continued efforts to interrogate the biology driving the growth of cancers will lead to benefits for trial design and drug development. Knowledge of the mechanisms driving cancer growth will present new potential drug targets, whilst knowledge of the ways in which cancers become resistant to treatments will enable rational combination drug trials to be designed. Trials need to focus on investigating better drugs rather than using “me-too” approaches to drug development. Trials of drugs with a strong preclinical basis should be prioritised for development, particularly those with reliable, robust predictive biomarkers to optimise patient selection. Drugs should be selected that potentially have a better therapeutic window, greater target specificity, decreased toxicity, or superior PK characteristics, including dose linearity, compared with existing agents, and obtaining biologically relevant dose levels in preclinical data should be a priority. Increased focus should be placed on novel first-in-class drugs that have the potential to have a greater impact.

Clinical trial design will also continue to evolve. There will continue to be increased patient enrichment and selection for early phase trials, given the drive towards personalised medicine, which is dependent on the development of accurate, reliable and reproducible biomarker assays. Due to the importance of biomarkers, the need for tissue acquisition in phase I trials will remain paramount. NGS techniques are continuing to become more cost-effective, sophisticated and time-efficient, factors that will drive their continued use in phase I trials. The infrastructure required to rapidly and efficiently obtain and test fresh biopsies or archival tissue for molecular screening will become increasingly important for early phase trial units. The potential benefits of NGS in selection for phase I trials, however, still remains unclear. Personalised medicine continues to generate significant attention in oncology. In the future this is likely to influence not only the choice of therapy and the timing of its use, but also the dose a patient receives. Increased awareness of the role of pharmacogenomics and other patient-specific factors is likely to influence dose selection. “N of 1” trials have been used in non-oncology drug development. In these trials, patients can be randomly assigned to different agents in a sequential order, with a wash-out in between, thus acting as their own control. This is a particularly interesting approach for patients with very rare molecular aberrations. The WINTHER trial (NCT01856296) is a modified “N of 1” trial design in which over 200 patients with different cancer types will have the PFS following treatment with a therapy chosen using advanced profiling techniques compared with the PFS on the regimen used immediately prior to trial enrolment [70]. Case reports of “N of 1” treatment are regularly published, but there is a bias towards positive results, leading Schilsky to propose a national registry of off-label targeted drug use to capture the outcomes of all such trials [71]. The Targeted Agent and Profiling Utilization Registry (TAPUR) Study (NCT02693535) is attempting to further explore the role of off-label use of FDA-approved targeted therapies. In the TAPUR study, the American Society of Clinical Oncologists’ first ever trial, patients are allocated to one of 15 arms, based on the presence of actionable mutations, with each arm assessing off-label indications of an FDA-approved targeted agent.

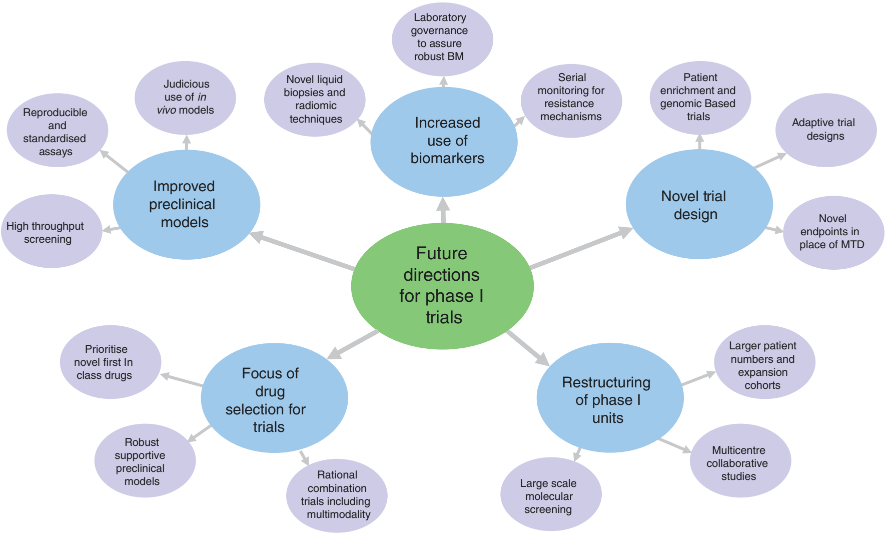

Future directions for phase I trials
Improved preclinical models
Reproducible and standardised assays
Judicious use of *in vivo* models
High throughput screening
Increased use of biomarkers
Novel liquid biopsies and radiomic techniques
Laboratory governance to assure robust BM
Serial monitoring for resistance mechanisms
Novel trial design
Patient enrichment and genomic Based trials
Adaptive trial designs
Novel endpoints in place of MTD
Focus of drug selection for trials
Prioritise novel first In class drugs
Robust supportive preclinical models
Rational combination trials including multimodality
Restructuring of phase I units
Larger patient numbers and expansion cohorts
Multicentre collaborative studies
Large scale molecular screening

Fig. 10.3 Future directions for early phase trials. The key areas within the design and implementation of phase I trials that are evolving to allow early phase trials to deliver improved outcomes are highlighted. This includes improved preclinical models, to ensure trials are built on a sound scientific foundation. Novel trial designs should be increasingly used to enable more efficient investigation of new agents. The increased use of biomarkers within phase I trials will allow focused identification of patients for recruitment to trials. Drug selection is key for early phase trials and drugs with novel targets or mechanisms of action; better target engagement or improved toxicity profiles should be prioritised. Phase I units will need to restructure to manage larger trials and multicentre collaborations and to support large-scale molecular screening programmes. *BM* biomarkers, *MTD* maximum tolerated dose

In the future trial designs will need to have an increased focus on monitoring the engagement of drug targets and the biological effects of a drug, rather than utilising toxicity alone as the primary endpoint. This will necessitate the inclusion of more comprehensive PK and PD studies within phase I trials. Given the limitations of single-agent treatments, combination trials of MTAs, immune oncology agents, chemotherapies and radiotherapy are likely to become increasingly common in the drive to improve patient outcomes. More sophisticated designs, including adaptive methods, are needed to meet the added challenges of these trials. Patient enrichment will continue to be important in early phase trials, with increased numbers of basket and umbrella trials being undertaken to test novel MTAs. Large phase 1b dose expansions are also likely to be increasingly common, given their potential to streamline and speed up drug development, particularly given the potential for FDA “break through therapy” designation. This will drive changes in phase I units, which will need to adapt to larger patient numbers, rapid patient recruitment and involvement in multicentre studies and large-scale molecular screening programmes (Fig. 10.3).

Conclusions

An ideal phase I trial would be efficiently performed with the minimum number of patients being exposed to sub-therapeutic or toxic doses, leading to accurate determination of an RP2D, a toxicity profile for an agent and a decision about its ongoing development. Sadly, the designs for such trials remain elusive, with clinical trials in the real world being beset by compromises, particularly given the tension between the need for rapid drug development due to cost and the need for generating optimal data and dose decisions. There has been an exponential increase in the number of oncological agents being trialled, but the approval rate for oncology drugs remains poor and the time for drug development is prolonged compared with the time for the development of drugs with non-oncological indications [72, 73]. Continuing to run the same style of clinical trials as we have in the past will not lead to improvements in the efficiency and cost-effectiveness of drug development. However, there is room for optimism, given the increasing prevalence of new technologies—such as high-throughput screening techniques and computational-based network platforms—to drive rational combinations, as well as adaptive trial designs to meet these challenges.

References

1. US FDA (Editor). Guidance for industry estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers. Silver Spring, MD: US FDA; 2005.
2. Janne PA, et al. AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. N Engl J Med. 2015;372(18):1689–99.
3. Storer BE. Design and analysis of phase I clinical trials. Biometrics. 1989;45(3):925–37.
4. Ji Y, et al. A modified toxicity probability interval method for dose-finding trials. Clin Trials. 2010;7(6):653–63.
5. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. Stat Med. 1998;17(10):1103–20.
6. Adaptive Designs Working Group of the MRC Network of Hubs For Trials Methodology Research “a quick guide why not to use A+B designs”. http://www.methodologyhubs.mrc.ac.uk/files/6814/6253/2385/A_quick_guide_why_not_to_use_AB_designs.pdf.
7. Tighiouart, M. and A. Rogatko, Dose finding with escalation with overdose control (EWOC) in cancer clinical trials. Stat Sci. 2010;(2):217–226.
8. Ji Y, Wang SJ. Modified toxicity probability interval design: a safer and more reliable method than the 3 + 3 design for practical phase I trials. J Clin Oncol. 2013;31(14):1785–91.
9. Yin G, Zheng S, Xu J. Fractional dose-finding methods with late-onset toxicity in phase I clinical trials. J Biopharm Stat. 2013;23(4):856–70.
10. EMEA (Editor). Guideline on strategies to identify and mitigate risks for first-in-human and early clinical trials with investigational medicinal products. London: EMEA; 2016.
11. US FDA (Editor). Guidance for industry S9 nonclinical evaluation for anticancer pharmaceuticals. Silver Spring, MD: US FDA; 2010.
12. Le Tourneau C, et al. Choice of starting dose for molecularly targeted agents evaluated in first-in-human phase I cancer clinical trials. J Clin Oncol. 2010;28(8):1401–7.
13. Hansen AR, et al. Choice of starting dose for biopharmaceuticals in first-in-human phase I cancer clinical trials. Oncologist. 2015;20(6):653–9.
14. Wong KM, Capasso A, Eckhardt SG. The changing landscape of phase I trials in oncology. Nat Rev Clin Oncol. 2016;13(2):106–17.

15. Elisei R, et al. Cabozantinib in progressive medullary thyroid cancer. J Clin Oncol. 2013;31(29):3639–46.
16. Sachs JR, et al. Optimal dosing for targeted therapies in oncology: drug development cases leading by example. Clin Cancer Res. 2016;22(6):1318–24.
17. Hughes A, et al. Development and evaluation of a new technological way of engaging patients and enhancing understanding of drug tolerability in early clinical development: PROACT. Adv Ther. 2016;33(6):1012–24.
18. Landers D. Technology – making trials simpler. Early phase workshop delivery of early phase oncology trials: how can we excel? London: CRUK Center for Drug Development; 2017.
19. Yeo WL, et al. Erlotinib at a dose of 25 mg daily for non-small cell lung cancers with EGFR mutations. J Thorac Oncol. 2010;5(7):1048–53.
20. Binder D, et al. Erlotinib in patients with advanced non-small-cell lung cancer: impact of dose reductions and a novel surrogate marker. Med Oncol. 2012;29(1):193–8.
21. Tol J, et al. Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. N Engl J Med. 2009;360(6):563–72.
22. Sandler A, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med. 2006;355(24):2542–50.
23. Gianni L, et al. AVEREL: a randomized phase III Trial evaluating bevacizumab in combination with docetaxel and trastuzumab as first-line therapy for HER2-positive locally recurrent/metastatic breast cancer. J Clin Oncol. 2013;31(14):1719–25.
24. Rini BI, et al. Phase III trial of bevacizumab plus interferon alfa versus interferon alfa monotherapy in patients with metastatic renal cell carcinoma: final results of CALGB 90206. J Clin Oncol. 2010;28(13):2137–43.
25. Ang JE, Kaye S, Banerji U. Tissue-based approaches to study pharmacodynamic endpoints in early phase oncology clinical trials. Curr Drug Targets. 2012;13(12):1525–34.
26. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92.
27. Cook N, et al. Early phase clinical trials to identify optimal dosing and safety. Mol Oncol. 2015;9(5):997–1007.
28. Hunsberger S, et al. Dose escalation trial designs based on a molecularly targeted endpoint. Stat Med. 2005;24(14):2171–81.
29. Mandrekar SJ, Cui Y, Sargent DJ. An adaptive phase I design for identifying a biologically optimal dose for dual agent drug combinations. Stat Med. 2007;26(11):2317–30.
30. Polley MY, Cheung YK. Two-stage designs for dose-finding trials with a biologic endpoint using stepwise tests. Biometrics. 2008;64(1):232–41.
31. Goulart BH, et al. Trends in the use and role of biomarkers in phase I oncology trials. Clin Cancer Res. 2007;13(22 Pt 1):6719–26.
32. Jardim DL, et al. Predictive value of phase I trials for safety in later trials and final approved dose: analysis of 61 approved cancer drugs. Clin Cancer Res. 2014;20(2):281–8.
33. Mok TS, et al. Osimertinib or platinum-pemetrexed in EGFR T790M-positive lung cancer. N Engl J Med. 2017;376(7):629–40.
34. Von Hoff DD, et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. J Clin Oncol. 2010;28(33):4877–83.
35. Manji A, et al. Evolution of clinical trial design in early drug development: systematic review of expansion cohort use in single-agent phase I cancer trials. J Clin Oncol. 2013;31(33):4260–7.
36. Sherman RE, et al. Expediting drug development—the FDA's new “breakthrough therapy” designation. N Engl J Med. 2013;369(20):1877–80.
37. Kramer DB, Kesselheim AS. User fees and beyond—the FDA Safety and Innovation Act of 2012. N Engl J Med. 2012;367(14):1277–9.
38. Kesselheim AS, Darrow JJ. FDA designations for therapeutics and their impact on drug development and regulatory review outcomes. Clin Pharmacol Ther. 2015;97(1):29–36.
39. Hamid O, et al. Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. N Engl J Med. 2013;369(2):134–44.

40. Trippa L, Alexander BM. Bayesian Baskets: a novel design for biomarker-based clinical trials. J Clin Oncol. 2017;35:PMID: 28045624.
41. Sleijfer S, Bogaerts J, Siu LL. Designing transformative clinical trials in the cancer genome era. J Clin Oncol. 2013;31(15):1834–41.
42. Project GENIE goes public. Cancer Discov. 2017;7(2):118.
43. Hyman DM, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. N Engl J Med. 2015;373(8):726–36.
44. McNeil C. NCI-MATCH launch highlights new trial design in precision-medicine era. J Natl Cancer Inst. 2015;107(7):pii: djv193.
45. Herbst RS, et al. Lung master protocol (lung-MAP)–a biomarker-driven protocol for accelerating development of therapies for squamous cell lung cancer: SWOG S1400. Clin Cancer Res. 2015;21(7):1514–24.
46. Chabner BA, Roberts TG Jr. Timeline: chemotherapy and the war on cancer. Nat Rev Cancer. 2005;5(1):65–72.
47. Audeh MW, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. Lancet. 2010;376(9737):245–51.
48. Gao S, et al. Applications of RNA interference high-throughput screening technology in cancer biology and virology. Protein Cell. 2014;5(11):805–15.
49. Day D, Siu LL. Approaches to modernize the combination drug development paradigm. Genome Med. 2016;8(1):115.
50. Gao H, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med. 2015;21(11):1318–25.
51. Mathews Griner LA, et al. High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. Proc Natl Acad Sci U S A. 2014;111(6):2349–54.
52. Iorns E, et al. Utilizing RNA interference to enhance cancer drug discovery. Nat Rev Drug Discov. 2007;6(7):556–68.
53. Keith CT, Borisy AA, Stockwell BR. Multicomponent therapeutics for networked systems. Nat Rev Drug Discov. 2005;4(1):71–8.
54. Paller CJ, et al. Design of phase I combination trials: recommendations of the Clinical Trial Design Task Force of the NCI Investigational Drug Steering Committee. Clin Cancer Res. 2014;20(16):4210–7.
55. Thall PF, et al. Dose-finding with two agents in phase I oncology trials. Biometrics. 2003;59(3):487–96.
56. Huang X, et al. A parallel phase I/II clinical trial design for combination therapies. Biometrics. 2007;63(2):429–36.
57. Yuan Y, Yin G. Sequential continual reassessment method for two-dimensional dose finding. Stat Med. 2008;27(27):5664–78.
58. Yin G, Li Y, Ji Y. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. Biometrics. 2006;62(3):777–84.
59. Polley MY. Practical modifications to the time-to-event continual reassessment method for phase I cancer trials with fast patient accrual and late-onset toxicities. Stat Med. 2011;30(17):2130–43.
60. Riviere MK, et al. Designs of drug-combination phase I trials in oncology: a systematic review of the literature. Ann Oncol. 2015;26(4):669–74.
61. Osborne CK, Kitten L, Arteaga CL. Antagonism of chemotherapy-induced cytotoxicity for human breast cancer cells by antiestrogens. J Clin Oncol. 1989;7(6):710–7.
62. Yap TA, Omlin A, de Bono JS. Development of therapeutic combinations targeting major cancer signaling pathways. J Clin Oncol. 2013;31(12):1592–605.
63. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. Nat Biotechnol. 2012;30(7):679–92.
64. Blomme EA, Will Y. Toxicology strategies for drug discovery: present and future. Chem Res Toxicol. 2016;29(4):473–504.

65. Bedard PL, et al. A phase Ib dose-escalation study of the oral pan-PI3K inhibitor buparlisib (BKM120) in combination with the oral MEK1/2 inhibitor trametinib (GSK1120212) in patients with selected advanced solid tumors. Clin Cancer Res. 2015;21(4):730–8.
66. Postel-Vinay S, et al. Challenges of phase 1 clinical trials evaluating immune checkpoint-targeted antibodies. Ann Oncol. 2016;27(2):214–24.
67. de Jonge MJ, et al. Phase I and pharmacokinetic study of pazopanib and lapatinib combination therapy in patients with advanced solid tumors. Investig New Drugs. 2013;31(3):751–9.
68. Reardon DA, et al. A phase I/II trial of pazopanib in combination with lapatinib in adult patients with relapsed malignant glioma. Clin Cancer Res. 2013;19(4):900–8.
69. Larkin J, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. N Engl J Med. 2015;373(1):23–34.
70. Rodon J, et al. Challenges in initiating and conducting personalized cancer therapy trials: perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. Ann Oncol. 2015;26(8):1791–8.
71. Schilsky RL. Implementing personalized cancer care. Nat Rev Clin Oncol. 2014;11(7):432–8.
72. Hay M, et al. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32(1):40–51.
73. DiMasi JA, Grabowski HG. Economics of new oncology drug development. J Clin Oncol. 2007;25(2):209–16.
74. Le Tourneau C, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol. 2015;16(13):1324–34.
75. Kim ES, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1(1):44–53.
76. Middleton G, et al. The National Lung Matrix Trial: translating the biology of stratification in advanced non-small-cell lung cancer. Ann Oncol. 2015;26(12):2464–9.
77. Topalian SL, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med. 2012;366(26):2443–54.

The Many Different Designs of Phase II Trials in Oncology

11

Rachel P. Riechelmann, Raphael L. C. Araújo, and Axel Hinke

11.1 Introduction

Phase II clinical trials comprise a crucial step in the treatment development process in oncology. These studies aim to evaluate preliminary signals of efficacy and safety of a given treatment in a specific population, and at the same time, to screen out inactive pharmacological agents. After a phase I trial determines the tolerable dose of drug or drug combination that can be safely administered in humans, a phase II trial assesses whether the experimental therapy may work. These trials generally do not establish a new standard of care, but rather provide initial efficacy data that can be—or not—validated in a subsequent confirmatory phase III clinical trial. In other words, phase II trials are screening instruments to determine whether a drug should be tested further in large phase III clinical trials.

To estimate the premature signals of efficacy of a new treatment many different phase II designs have been proposed, taking into account numerous facets: tumor and patient characteristics, cancer frequency in the population, the anticipated anticancer effects of the experimental agent(s), patients' expected outcomes with available therapies, and budget. Here we discuss the most commonly utilized phase II trial designs to evaluate the efficacy of new cancer therapies, as well as the methodological implications and advantages and disadvantages of each design.

R. P. Riechelmann, M.D., Ph.D. (✉)
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil
e-mail: rachel.riechelmann@hc.fm.usp.br

R. L. C. Araújo, M.D., Ph.D.
Department of Upper Gastrointestinal and Hepato-Pancreato-Biliary Surgery, Barretos Cancer Hospital, Barretos, SP, Brazil

A. Hinke, Ph.D.
CCRC, Düsseldorf, Germany

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_11

11.2 General Aspects of Phase II Trials in Oncology

Phase II cancer trials are larger than phase I studies but smaller than phase III trials. They often enroll between 40 and 100 patients and are completed in 2–3 years. Therefore they offer quicker results than phase III studies and are thus considered "fast" trials. One of the reasons for their being fast is that they predominantly use surrogate endpoints as measures of drug efficacy. The great majority of these trials are entirely or partially funded by for-profit companies [1, 2], and more and more, they recruit patients with molecularly homogeneous tumor features and restricted eligibility criteria. Because of the known cancer heterogeneity, testing a drug in an unselected patient population may lead to false-negative results and thus hamper its further development into a phase III trial. In contrast, the antineoplastic effects of a drug can be more adequately evaluated if the study population is homogeneous enough in terms of clinical and tumor characteristics. For example, a trial of an anti-HER2 agent in patients with gastric cancer should be done in patients whose tumors overexpress HER2.

To properly determine drug efficacy in phase II trials investigators have to predefine endpoints. Of note, the study objective is totally distinct from the study endpoint, i.e., the objective relates to what we want to achieve, and the endpoint relates to *how* we aim to assess a positive outcome. In this regard, the main objective of a phase II trial is preliminary evidence of efficacy, while the endpoints define what efficacy means in the context of the trial. Frequently used endpoints in oncology include the response rate (RR) for cytotoxic agents, such as chemotherapy, and time-to-event endpoints such as progression-free survival (PFS), often used to measure the antitumor activity of cytostatic drugs [1]. The disease control rate (DCR) and time to progression (TTP) have also been utilized, while patient-reported outcomes, such as pain, fatigue, nausea, and/or elements of quality of life, are less commonly employed. The RR often relies on the RECIST (response evaluation criteria in solid tumors) [3], which generally define the proportion of patients whose tumors' largest diameters shrink by at least 30% on imaging tests in comparison with the baseline values, usually measured after 8–12 weeks from the initial treatment. The DCR is also a categorical variable—it expresses the proportion of patents without tumor progression, usually determined by RECIST, at a pre-defined time point. PFS is defined as the time between the first day of treatment or the date of randomization until the date of disease progression or death; patients lost to follow-up or patients who did not reach the event at the time of analysis are censored. TTP is similar to PFS, but death is treated as a censored event, i.e., if a patient dies on study, without formal proof of considerable tumor growth, the death is not assumed to be due to progression. TTP and PFS may present similar findings in trials of indolent tumors treated with gentle, less toxic agents, while PFS is certainly preferred in trials of toxic drugs, because toxicity-induced deaths are considered as an unfavorable event, similar to progression, rather than being regarded as a censored event, in the intention-to-treat analysis. For example, the efficacy of somatostatin analogues, which are nontoxic agents, in well differentiated neuroendocrine tumors can be assessed by TTP, while a trial of cytotoxic chemotherapy for advanced colorectal cancer should utilize PFS instead of TTP as a measure of drug efficacy. Importantly, all these endpoints are considered surrogates of the true endpoints of patient benefit, which mostly focuses on overall survival and quality of life. Phase II trials mostly use surrogate endpoints, while phase III trials regularly evaluate true endpoints to establish a new standard of care [4].

In the past decades large financial resources, provided by pharmaceutical companies and governments/not-for-profit organizations, have been invested in cancer research. Consequently the number of new cancer drugs has increased significantly, requiring an expedited and "cost-effective" drug development process to evaluate drug efficacy. In this scenario, phase II trials, with their different designs, play a central role in determining "go *versus* no-go" for a given therapy being tested in phase III trials.

11.3 How to Plan a Phase II Clinical Trial?

When planning a phase II trial, just as in any other type of study, an investigator first has to formulate the research question. This initial question may cover a broad topic, but it has to be further detailed so that the trial is designed to answer a specific issue. Once the question is defined, the following steps have to be determined a priori: study design and number of study arms, patient population, trial endpoints (primary and secondary) and their assessments, treatment and follow-up schedules, planning of statistical analysis and sample size calculation, and last, but not least, the study budget.

The study population is determined by the eligibility criteria. These criteria comprise an extensive list of inclusion and exclusion criteria that subjects have to comply with. Such criteria should provide detailed descriptions of, e.g., the participants' clinical *performance status*; tumor histological—and often molecular—type; cancer stage; presence of brain metastases; organ functions, including bone marrow and hepatic functions; information on allowed prior therapies; permitted comorbid illnesses; and concomitant medications (to prevent undesirable drug-drug interactions).

Determination of the eligibility criteria should consider an adequate balance between internal and external validity. Internal validity reflects the strength of causality, i.e., whether the drug actually caused an effect on the tumor. In other words, internal validity informs on how well a study was conducted; thus, the higher the internal validity, the better the quality of the trial as a scientific experiment. To prove causality, the researcher must maximize the control of study variables and subjects. This is done by enhancing the homogeneity of the features of the study patients and interventions, such as by having strict eligibility criteria and uniform frequency of treatment administrations, imaging tests, and follow-up periods. However, if investigators control the experiment excessively; for example, by over-restricting the eligibility criteria, the study may not be completed due to a lack of eligible subjects or if it is completed, its results may not be applicable to the general population with the same cancer type and stage. This extent to which a trial result can be generalized to other similar populations is called external validity. Studies have shown that fewer than 5% of cancer patients are enrolled into clinical trials, and the selection process is far from random, e.g., with respect to the patients' average age [5]. So, in reality, the outcomes of most cancer-directed interventions delivered to community patients are overwhelmingly unknown. Large population databases, or non-interventional, observational studies, e.g., post-marketing phase IV studies, can help fill this gap and have tried to provide society with complementary evidence on the outcomes of oncology treatments administered in 'real-life patients'.

The selection of the primary endpoint is also crucial for planning the trial design and calculating the sample size (discussed in Chap. 5). As noted above, RR and PFS are the most frequently selected primary endpoints, with RR being preferred in trials where tumor shrinkage is expected, and PFS preferred in trials of cytostatic drugs that offer tumor stabilization [6]. Secondary endpoints are measures of treatment efficacy and safety that accompany the primary endpoint. Secondary endpoints are never used for sample size calculation, and thus their results might be insufficiently powered to achieve reliable evidence or drive the study primary conclusion. Yet their results are important because they complement the primary result and/or may generate hypotheses to be tested in further trials. Common secondary endpoints of phase II trials are toxicity profiles and their corollaries (dose intensity, rate of treatment interruptions/discontinuation, drug exposure, pharmacokinetics), overall survival and survival rate at specific time points, RR determined by methods other than RECIST, PFS, and sometimes, patient-reported outcomes. Correlative or translational studies have become increasingly usual in oncology trials. They look for "biological endpoints" that try to evaluate the prognostic and predictive effects of tumor molecular markers that are identified in tumor biopsies. A prognostic marker is associated with disease outcome independently of treatment, while a predictive marker is correlated with the response to a given drug. For example: mutations in the V600 BRAF oncogene are prognostic in metastatic colorectal cancer [7] and apparently predict resistance to anti-epidermal growth factor (EGFR) agents in RAS wild-type tumors [8]. Pharmacodynamic endpoints are mostly utilized in phase I trials to test the biological activity of a targeted agent.

The determination of treatment schedules and drug doses must be based on data derived from phase I trials (see Chap. 10). Pre-medications and supportive therapy permitted or recommended to be taken during the trial, as well as guidelines for dose modifications/interruptions resulting from toxicities, have to be detailed in the phase II protocol as well.

The statistical analyses must be planned before starting the trial. This includes the definition of study endpoints, how variables will be defined (continuous versus categorical), whether statistical tests will be used for inferences, and if so, which tests and significance level will be considered, and of course, sample size computation. Usually, the key analyses should be undertaken following the intention-to-treat principle.

All these steps discussed here should be followed when developing the phase II protocol *and* they should be reported when trial results are published. The reporting of all these parameters is very important so that readers can properly interpret the study results. Unfortunately, the reporting of these parameters is frequently not done. A cross-sectional study of 125 phase II trials in oncology showed that 27% did not clearly report the primary endpoint [1]. In another survey, of 295 phase II cancer trials, the statistical design could not be appraised in 19.7% [9].

11.3.1 Different Designs of Phase II Trials

11.3.1.1 Single-Arm Phase II Trials

The first phase II design to be widely used in clinical cancer research was the one-stage design. In this design, usually 40–50 patients were treated to assess whether at least 20% of the patients experienced RR with an experimental cytotoxic agent. It is not known where this "rule" came from or when it was first used, but it is likely that during the 1950s doctors just started using this number of patients in single-arm phase II trials, making this a common practice.

With the evolution of clinical cancer research, the scientific community started to ask whether it would be possible to make phase II trial results even more quickly available and thus treat fewer patients with ineffective drugs. Instead of treating 40 patients to determine whether a cancer-directed agent did not work, could a trial give that answer reliably if it treated only 15–20 patients? This question set the basis for the designs of phase II trials with two or more stages. Gehan [10] envisioned a two-stage design in which a trial is terminated early if any favorable response is detected. Fleming [11] proposed a multi-stage design that permitted early stopping because of futility or efficacy, and Simon [12] developed a two-stage optimized design that minimized the number of enrolled patients. The most common staged design, the two-stage phase II trial, was developed to screen for drug efficacy (measured by RR) with a minimum number of patients. These trials commonly enroll a first stage with a comparatively small patient number; if no or an insufficient number of responses are seen, the probability of success is estimated to be very low and the trial is terminated. In contrast, if a pre-defined minimum number of responses is observed, a second stage of enrollment is carried out until a pre-defined total number of patients is achieved. The sample size is based on RR probabilities considered to be "futile" or "promising" (for further development). Nowadays it is more common that patients enrolled in two-stage phase II trials receive standard therapy plus an experimental treatment. In this case, even if there are many responses (which can be attributable to the standard regimen), the trial may still be stopped early.

The sample size calculation assumption for phase II single-arm trials relies on historical data, i.e., the estimated gain with the experimental therapy over the standard treatment is based on information from publications or on institutional retrospective data. This automatically leads to a selection bias, because the phase II results, not being controlled by randomization, can be over- or underestimated against the results of the standard therapy.

The argument against single-arm phase II trials is that they might have low power to detect small but clinically relevant differences, which may lead to early termination of trials with drugs that could show activity in a larger sample. In contrast, one might argue that highly effective therapies would show some activity, even in a small sample of patients. Additionally, two-stage designs are very useful when there are many drugs to be tested, a restrictive budget, and short timelines.

The general advantages of conducting single-arm phase II trials are: they are fast and cheap studies to screen for drug efficacy; it would certainly not be feasible to perform randomized phase II trials to initially evaluate efficacy for every single new antineoplastic agent. Disadvantages of single-arm phase II trials include the lack of a control arm, thus implying a high risk of selection bias that compromises external validity. Selection bias, in turn, may lead to false-positive or -negative results. Moreover, time-to-event endpoints, such as PFS and TTP, are difficult to estimate in phase II trials because of their great variability when compared with RR. This is because PFS measurement requires more frequent imaging tests and is deeply influenced by intervals between imaging tests [13]. In addition, event rates do not reflect any direct antineoplastic effect (even when there is no efficacy at all, some non-zero PFS rate will emerge). Consequently, single-arm phase II trials should not guide clinical practice, as promising and clinically relevant findings from these trials have to be confirmed by phase III trials.

11.3.1.2 Multiple-Arm Phase II Trials: Biomarker-Driven Trials

Biomarker-driven trials (BDTs) are phase II trials with multiple study arms, where patients are molecularly selected to receive directed therapy. Patients may be randomized or not, and if not randomized, they are allocated to study arms according to their molecular eligibility criteria. These trials have recently become popular with the observation that specific molecular alterations and altered pathways are encountered in different tumor types. These trials rely on the hypothesis that the presence of molecular alterations predicts treatment benefit from specific targeted therapies, regardless of cancer type. Thus, a drug that inhibits a mutation X may be tested in a BDT trial of multiple arms with patients whose tumors harbor mutation X.

One of the first BDT trials was the BATTLE (Biomarker-Integrated Approaches Of Targeted Therapy for Lung Cancer Elimination) trial [14]. In this trial molecularly selected patients with advanced non-small cell lung cancer whose disease had progressed after conventional chemotherapy were allocated to one of five biomarker-specific cohorts; within each cohort patients were randomized to one of four arms, generating a 20-arm trial. The BATTLE trial used a response adaptive randomization strategy, where the patients would have a higher chance of being randomized to the study arm with the greatest probability of response. Since that trial, different versions of BDTs have been conducted, as summarized below:

- Umbrella trials: In these BDT trials patients with specific cancers have their tumors profiled, and according to the molecular alterations encountered, they are allocated to particular targeted therapies—or randomized to targeted therapy versus standard of care. For example, in the Multicenter Study of Biomarker-Driven Therapy in Metastatic Colorectal Cancer, the MODUL trial (NCT02291289), patients with RAS wild-type treatment-naïve metastatic colorectal cancer with a

BRAF mutation will receive an oral BRAF inhibitor combined with an anti-EGFR monoclonal antibody and 5-fluorouracil (FU), while those with HER2 amplification will receive trastuzumab, pertuzumab, and capecitabine. In this trial, new biomarker-driven cohorts can be added based on new relevant scientific information.

- Basket trials: These BDTs have a unique genomic platform, often next-generating sequencing, where tumors are profiled and patients with different tumors are assigned to a biomarker-specific single arm. An example is the American National Cancer Institute-sponsored MATCH trial, which opened for accrual in August 2015 (https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match). The MATCH trial plans to screen nearly 3000 patients to enroll approximately 1000 into 24 marker-specific independent single-arm parallel studies aimed at matching tumor mutations regardless of the tumor primary site.
- Marker strategy design: These trials compare a biomarker-specific group treated with a targeted therapy versus such a group receiving a non-targeted therapy. Variations of this design include randomization of patients without the biomarker to also receive the targeted therapy [15]. This design is useful when investigators are not certain about the predictive relevance of the biomarker.

The advantages of BDTs lie in their smart approach to tackle tumor-specific driver mutations and to assess response in relatively small samples [15]. BDTs are also capable of validating predictive biomarkers. The disadvantages of BDTs relate to their high cost, potentially insufficient knowledge about a biomarker (which may render negative results), requirement for high-quality tumor tissues and timely results of molecular analyses, and the need for skillful personnel to guarantee accurate molecular results. Particular challenges in designing BDTs include the inherent variability of study patients. For example, in basket trials, a mutation may not be predictive of benefit from a targeted agent in all tumor types harboring that mutation, and in umbrella trials, different molecular profiles might be associated with a distinct prognosis; also, knowing the frequency of mutations in the study population is crucial for estimating the screening sample size and its associated budget.

11.3.2 Randomized Phase II Trials

Randomized phase II (RPh2) trials are comparative clinical trials that are also designed to preliminarily evaluate the efficacy of anticancer agents. They are not supposed to establish a new standard of care because they are not powered to detect differences in true oncology endpoints—overall survival or quality of life. These trials resemble two single-arm phase II trials, with the difference that the study arm allocation is determined by randomization. And that is the main strength of RPh2 trials: the capability to control for selection bias, stage migration, and changes in medical practice over time and/or by geographical region and/or by type of participating institutions, factors that are often encountered in single-arm trials. The weaknesses of RPh2 trials lie in their inability to identify small differences across study arms, owing to lack of power.

RPh2 trials have been devised as a way to expedite drug development in an era with increasing numbers of drugs to be tested. These trials consist of two or multiple arms whose different results are not supposed to be statistically compared but are ranked in order to pick the "best" result to test in phase III trials, regardless of the magnitude of the difference [13]. This is because to perform mathematical inferences would require large sample sizes, which, in turn, would delay rather than accelerate trial results. Hence, RPh2 trials were not envisioned to provide statistical comparisons across the study arms but to simply describe the outcomes. Yet the oncology research community has practiced differently. A survey of 107 RPh2 trials published in a decade showed that either *p* values or confidence intervals, as means of formal statistical comparisons, were reported in nearly 89% of these trials [2]. In the event that statistical evaluations are performed, investigators must state that such inferences are purely exploratory and should not guide clinical practice. RPh2 trials cannot substitute for a phase III trial. Indeed, regulatory agencies rarely accept results from RPh2 trials to approve the commercialization of new cancer-directed drugs, except in rare instances of exceptional drug activity in orphan diseases.

RPh2 trials are more complex to design and conduct than single-arm phase II studies, as they recruit larger numbers of patients. These trials often have two study arms, enroll approximately 50–200 patients, utilize RR or event rates at specific time points as the most frequent endpoints, and often receive industry funding [2]. Their designs follow those of single-arm phase II trials, i.e., each trial arm is planned as a one- or two-stage phase II trial. The particularities of the most commonly utilized RPh2 designs are discussed in detail below.

11.3.2.1 Pick the Winner

The "pick the winner" design was the first widely elected one to be used in RPh2 trials. As the name suggests, such trials report outcomes by ranking them from the best—or winner—to the worst result of the primary endpoint. These trials may test two or more therapies in terms of monotherapy, monotherapy versus combination therapy, or different combination regimes or associations with radiation etc., with all study arms examining an experimental intervention. Given that statistical comparisons between or across trial endpoints are not performed, deciding which arm reports the winner result may not be easy. The investigator's discretion is the key to picking the winner therapy to be further tested in phase III trials. Of note, the pick the winner design is particularly useful for screening out inactive treatments. For example, in a three-arm RPh2 trial, the RR was 25% in arm A, 28% in arm B, and 12% in arm C. Assuming the standard therapy would offer an RR of 10%, arm C is clearly inferior to arms A and B and this intervention might be discarded; on the other hand, the decision to pick either arm A or B should consider the toxicity profiles of the different therapeutic regimes, because, in terms of activity, arms A and B suggest similar efficacy.

11.3.2.2 Randomized Phase II Trials with Standard Control

RPh2 trials may also elect one of the study arms as the standard treatment, functioning as a means of internal control. Such trials often utilize surrogate endpoints such as time-to-event endpoints or RR, and their sample size is calculated as for two (or more) independent single-arm phase II trials, which are controlled by randomization. In such a case, the standard arm is called the "calibration" arm, which aims to control for selection bias and to verify the historical assumption of the standard therapy. RPh2 trials in which one of the study arms is the standard treatment are not supposed to substitute for a formal phase III trial, but rather to have a control arm that serves as a comparator for the experimental intervention. However, RPh2 controlled trials may continue as a phase III trial if promising results are observed, as discussed in Sect. 11.3.2.4.

In some cases, RPh2 controlled trials are also designed to look for large differences between the study arms, often larger than those used when computing the sample size for a phase III trial [16]. This type of RPh2 trial allows for more "relaxed" type-I error levels ($p = 0.1$ or 0.2, even one-sided) and/or puts overoptimistic effect sizes (e.g., hazard ratio = 0.6) into the calculation, in order to achieve "phase II type" sample sizes. Finally, RPh2 trials with standard control can be designed to use sensitive surrogate markers as the primary endpoint. In such cases, formal statistical comparisons can be made and standard sample size calculation procedures are similar to those for phase III trials.

RPh2 controlled trials are especially advantageous when the outcome of the standard therapy is unknown or not accurately described in the literature, i.e., the outcome has significant variability. For example, when planning an RPh2 trial to test the PFS provided by a new second-line therapy for patients with adenocarcinoma of the pancreas previously treated with the FOLFIRINOX regimen, investigators need to know what the anticipated PFS of the standard second-line therapy is in order to estimate the gain from the investigational drug. The problem is that currently there is no standard second-line therapy in this setting and most patients are treated with gemcitabine, based on heterogeneous retrospective data [17]. Therefore, in this case, having gemcitabine monotherapy as a comparator arm facilitates the interpretation of the results of the experimental therapy. The trial could simply describe the PFS provided by gemcitabine and the experimental therapy, or the investigators could power the trial so that the investigational arm would need to double the anticipated PFS offered by gemcitabine alone. In instances of trials of cytostatic drugs to treat indolent neoplasms, where tumor stabilization may take place without treatment, placebo exclusive or best supportive care can be used as the internal control when there is no effective standard therapy.

11.3.2.3 Randomized Discontinuation Trial

The randomized discontinuation trial (RDT) is a unique type of RPh2 design. It was first proposed by Amery and Dony in 1975, in a trial of patients with angina pectoris [18]. The trial was designed with the aim being to decrease the number of patients being exposed to placebo and at the same time, to control the effect of placebo on study outcomes. This phase II design was first used in oncology in a trial of

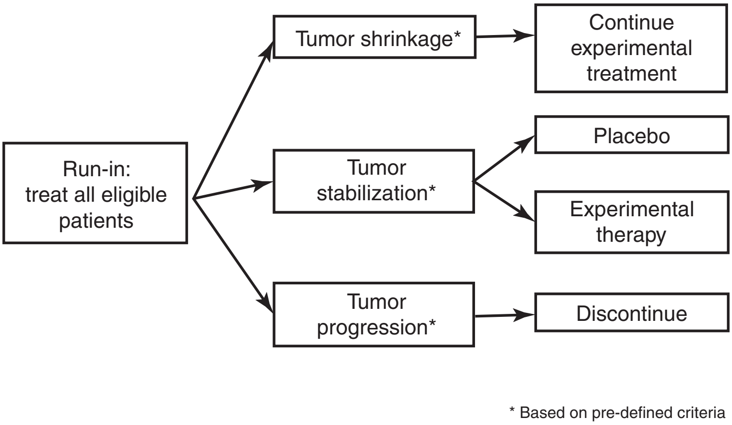


Fig. 11.1 Schema of randomized discontinuation trials

carboxyaminoimidazole for metastatic renal cell carcinoma [19], but the design became popular after an important trial of sorafenib for patients with advanced renal clear cell carcinoma [20]. In this design, a large number of patients are treated with the investigational agent for a pre-defined period, during which imaging tests are performed to measure drug activity. For patients whose tumors respond, the study drug is maintained until tumor progression; for patients who experience tumor progression, treatment is halted; while those who present tumor stabilization are randomized to either placebo or the experimental treatment (Fig. 11.1). Differently from most RPh2 trials, RDTs are designed for statistical comparisons between the randomized groups. In cases of indolent tumors, such as well-differentiated neuroendocrine tumors and renal clear cell carcinoma, formal "tumor control" may be achieved without any treatment, as a biological feature of the neoplasm. To measure drug efficacy in this setting, the group of patients with stable disease are randomized in a double-blind fashion to continue on the study drug or to be switched to placebo. In the aforementioned sorafenib trial, 202 patients were treated with sorafenib 400 mg orally twice daily for 12 weeks, at which point imaging tests were performed. Patients with growth in bidimensional tumor measurements that was equal to or higher than 25% of baseline were deemed as having progression and were discontinued from the trial ($N = 64$); those with tumor shrinkage of at least 25% were considered responders and were kept on trial until progression or intolerance ($N = 73$). Patients with stable disease ($N = 65$), defined by changes in tumor measurements from baseline of less than 25%, were randomized to sorafenib or matched-placebo in a double-blind manner [20]. The study was positive for its primary endpoint of the proportion of randomized patients who were progression-free at 24 weeks from the start of sorafenib: 50% for sorafenib-treated patients versus 18% for the placebo group ($p = 0.0077$) [20]. Of note, 18% of patients in the placebo group had tumor stabilization after 6 months; this finding highlights the importance of randomization in patients with indolent tumors treated with cytostatic agents, to avoid overrated interpretations or false-positive results.

The advantages of RDTs include minimization of placebo, because not all patients are randomized upfront; high internal validity and, consequently, greater statistical power owing to a more biologically homogeneous/enriched population; and their use as a tool to specifically measure the cytostatic effects of antineoplastic therapies in slow-growing tumors. Disadvantages of these trials are their complex resource- and time-consuming logistics. RDTs should not be conducted in a setting of an expected low rate of tumor stabilization, because this would require a prohibitively large sample. For example, if only 30% of patients are expected to be sensitive to treatment, defined by a 30% decrease in the tumor growth rate, a sample of 1650 patients would be necessary [21]. There is also concern about real blinding, as patients randomized to placebo may know what they are taking because they might have experienced drug-induced adverse events during the run-in phase. This may lead to placebo patients being more easily deemed as having progression and crossing over to the treatment arm. Given that these trials generally test targeted agents, one could argue that in positive RDTs patients with stable disease who are randomized to the placebo arm may do worse simply because of drug withdrawal, suggesting a theoretical rebound effect. Additionally, with the emergence of numerous and effective targeted agents for molecularly selected patients, RDTs may become less useful, while upfront randomization is preferable [21]. However, these trials may still be valuable to test multitarget-kinase inhibitors in unselected patients with advanced solid tumors. Finally, it may be difficult to conduct a subsequent phase III trial after a "very" positive RDT owing to the ethical implications of randomizing patients to a known inferior treatment (placebo) and also because the upfront randomization might dilute the drug efficacy in the overall population. Therefore RDTs have to be carefully planned, taking into account any subsequent phase III trial.

11.3.2.4 Randomized Phase II/III Trials

Randomized phase II/III trials are two-stage trials [22, 23]. The goal of the phase II portion is to screen for efficacy of the experimental therapy against the control or standard arm. If the results are encouraging the study is carried forward and additional patients are enrolled into a confirmatory phase III trial. The phase II and phase III portions of randomized phase II/III trials use the same design, endpoints, and eligibility criteria. The phase II sample size is computed taking into account the anticipated differences between the study arms in respect to the phase II primary endpoint, which is often a surrogate endpoint of overall survival. To expand to a phase III, the sample size is recalculated to estimate gains in true efficacy endpoints. Unlike the "pick the winner" approach, here the phase II arms are compared statistically rather than being ranked according to the primary endpoint results [22]. Randomized phase II trials that stop because of futility before continuing into a phase III are called "non-adaptive", e.g., the ABC-02 trial discussed below. However, if a "winner" arm is selected among several experimental arms, or if investigators change some aspects of study design at the end of phase II, the randomized phase II/III trial is called "adaptive".

The advantages of this phase II/III design include the accelerated move into phase III when there are encouraging initial findings in the phase II part. Reducing the time lag between the completion of a phase II and the launch of a phase III trial saves administrative and financial resources. Another benefit is the possibility of stopping the trial because of futility before going into phase III, which certainly minimizes the number of patients being treated with futile drugs.

A good example of a randomized phase II/III cancer trial is the ABC-02 study [24]. In this trial, patients with advanced biliary cancer were randomized to receive first-line gemcitabine monotherapy as the control arm, or gemcitabine combined with cisplatin, until disease progression. The trial enrolled 86 patients in the phase II stage using PFS as the surrogate efficacy endpoint. Once the enrollment was completed, the trial was stopped to evaluate whether the results were positive. Because the PFS of the combination therapy was longer, the phase II portion was deemed positive and more patients were accrued so the study continued into a phase III trial. While the same eligibility criteria and treatment regimens were used in both phases, the primary endpoint of the phase III portion was changed to overall survival, and the final sample size was 400 patients.

It is important to highlight that the evaluation of futility in interim analyses of phase III trials is totally different from the evaluation of the results at the end of the phase II portion in a randomized phase II/III design. The stopping rules in the interim analyses of phase III trials determine that the trial be terminated if the experimental arm is *worse* than the control arm; in randomized phase II/III trials, similar results for the two arms at the end of phase II abrogate further development into phase III [23].

11.3.3 Adaptive Phase II Trials

All the designs described in detail in this chapter are based on the classical "frequentist" approach to probabilities, error margins, statistical tests, and confidence intervals. Several novel designs based on Bayesian techniques have been developed as alternatives to such designs. In adaptive phase II trials, the parameters of the study are continuously monitored and modified according to the results. This process includes changes in sample size, dosages of drugs, and types of treatments (with the addition and exclusion of treatment groups). Advantages of these designs include that there are no restrictions on the number of looks into the accumulating data, that they are more flexible with respect to early stopping, and that they allow for the combination of outcome variables (e.g., in combined phase I/II trials). Disadvantages of adaptive phase II trials include that they depend heavily on pre-specified assumptions (prior distribution) and are more complicated to implement, as the many consecutive (and computationally intensive) analyses require an immediate distribution of validated endpoint observations to the statistical center. This may prove to be difficult in multicenter, or even in multinational, trials. Although Bayesian designs are advantageous in special situations, e.g., in phase II studies incorporating dose-finding, the vast majority of phase II trials published nowadays are still performed based on the "classical" statistical paradigms.

11.4 Conclusions and Perspectives

Phase II trials represent a key step in the drug development process of new cancer-directed therapies. They are useful tools to screen for drug efficacy and to set the basis for the decision to move into phase III trials. There are many different designs for phase II trials and each one has unique advantages and disadvantages. Consequently, continuous efforts are underway to further refine phase II designs, with the aim being to identify treatment efficacy promptly and at the same time to decrease the number of cancer patients who might receive ineffective therapies.

References

1. Riechelmann RP, Dounaevskaia V, Krzyzanowska MK. Quality of reporting primary outcomes in phase II cancer trials. J Clin Oncol. 2008;26(33):5486–8.
2. Saad ED, Sasse EC, Borghesi G, et al. Formal statistical testing and inference in randomized phase II trials in medical oncology. Am J Clin Oncol. 2013;36(2):143–5.
3. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2017;45(2):228–47.
4. Schilsky RL. End points in cancer clinical trials and the drug approval process. Clin Cancer Res. 2002;8(4):935–8.
5. Cassileth BR. Clinical trials: time for action. J Clin Oncol. 2003;21(5):765–6.
6. Adjei AA, Christian M, Ivy P. Novel designs and end points for phase II clinical trials. Clin Cancer Res. 2009;15(6):1866–72.
7. Tran B, Kopetz S, Tie J, et al. Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. Cancer. 2011;117(20):4623–32.
8. Pietrantonio F, Petrelli F, Coinu A, et al. Predictive role of BRAF mutations in patients with advanced colorectal cancer receiving cetuximab and panitumumab: a meta-analysis. Eur J Cancer. 2017;51(5):587–94.
9. Mariani L, Marubini E. Content and quality of currently published phase II cancer trials. J Clin Oncol. 2000;18(2):429.
10. Gehan EA. The determinatio of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis. 1961;13:346–53.
11. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982;38(1):143–51.
12. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1–10.
13. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treat Rep. 1985;69(12):1375–81.
14. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1(1):44–53.
15. Renfro LA, An M-W, Mandrekar SJ. Precision oncology: a new era of cancer clinical trials. Cancer Lett. 2017;387:121–6.
16. Rubinstein LV, Korn EL, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol. 2005;23(28):7199–206.
17. da Rocha Lino A, Abrahão CM, Brandão RM, Gomes JR, Ferrian AM, Machado MC, Buzaid AC, Maluf FC, Peixoto RD. Role of gemcitabine as second-line therapy after progression on FOLFIRINOX in advanced pancreatic cancer: a retrospective analysis. J Gastrointest Oncol. 2015;6(5):511–5. https://doi.org/10.3978/j.issn.2078-6891.2015.041.

18. Amery W, Dony J. A clinical trial design avoiding undue placebo treatment. J Clin Pharmacol. 1975;15:674–9.
19. Stadler WM, Rosner G, Small E, et al. Successful implementation of the randomized discontinuation trial design: an application to the study of the putative antiangiogenic agent carboxyaminoimidazole in renal cell carcinoma—CALGB 69901. J Clin Oncol. 2005;23(16):3726–32.
20. Ratain MJ, Eisen T, Stadler WM, et al. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. J Clin Oncol. 2006;24(16):2505–12.
21. Freidlin B, Simon R. Evaluation of randomized discontinuation design. J Clin Oncol. 2005;23(22):5094–8.
22. Schaid DJ, Wieand SAM, Therneau TM. Optimal two-stage screening designs for survival comparisons. Biometrika. 1990;77(3):507–13.
23. Korn EL. Design issues in randomized phase II/III trials. J Clin Oncol. 2012;30:667–71.
24. Valle J, Wasan H, Palmer DH, et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. N Engl J Med. 2010;362(14):1273–81.

Superiority and Non-inferiority Phase III Oncology Trials

12

Everardo D. Saad

12.1 Introduction

Phase III trials are considered the gold-standard approach for evaluating therapeutic interventions in medicine. The hierarchical prominence of these trials stems from their capacity to allow for inferences about causal links between treatment and outcomes. Given their explicit objectives and statistical features, phase III trials are the ideal scenario for comparing two or more competing systemic therapies for cancer. Most phase III trials aim to demonstrate the superiority of a new treatment in comparison with control; however, phase III trials may also assess whether a more convenient, less toxic, or more affordable intervention is at least as efficacious as an existing standard of care, and these are called non-inferiority trials. Given the fact that several features are common to superiority and non-inferiority trials, most of the discussion in this chapter applies to both types of phase III trials. When necessary, issues that are exclusive to superiority trials will be highlighted, and the last section of the chapter will be dedicated to non-inferiority trials. Although this chapter focuses on selected methodological issues of phase III trials, the reader should note that additional issues, as well as ethical, administrative, and operational aspects, are also important and must be taken into account in the design, conduct, and analysis of these trials [1, 2].

E. D. Saad, M.D.
Dendrix Research, São Paulo, SP, Brazil

IDDI, Louvain-la-Neuve, Belgium
e-mail: everardo@dendrix.com.br

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_12

12.2 General Design Features of Phase III Trials

12.2.1 Position in the Spectrum of Validity

A general property of experimental research is that internal validity (i.e., the reliability of results) and external validity (i.e., their generalizability) tend to move in opposite directions. This gives rise to different attitudes toward clinical trials, one that prioritizes internal validity (the explanatory attitude), and one that places more emphasis on the generalizability of results (the pragmatic attitude) [3]. This difference in attitudes is reflected in several features of the design and conduct of phase III trials, some of which are summarized in Table 12.1. The explanatory attitude is epitomized by pivotal trials of novel agents, which are required for drug approval and seek to assess the *efficacy* of such agents in somewhat idealized conditions. Pragmatic trials, often conducted by academic groups or governmental agencies, seek to compare the *effectiveness* of interventions that are already available in clinical practice under less controlled conditions that are closer to "real life". In reality, most trials display elements of both attitudes, and tools have been developed to determine the relative position of a trial in the explanatory-pragmatic continuum [4]. More recently, the pragmatic side of the spectrum has been expanded by comparative-effectiveness research, which is distinct from traditional phase III trials in its attempt to generate evidence in the setting of usual clinical care, as opposed to research-intensive or academic environments [5]. However, a discussion of comparative-effectiveness trials is beyond the scope of this chapter.

12.2.2 Number of Arms

The vast majority of phase III trials have one experimental and one control arm. Two-arm trials are easier to interpret and are generally preferable [6, 7]. On occasion, two or more experimental arms may be tested, but the statistical design for

Table 12.1 Some key elements in the explanatory-pragmatic spectrum

Trial feature	Explanatory	Pragmatic
Usual sponsor	Industry	Academia, government
Typical setting	Pivotal trial of a novel agent	Comparison of two available agents or combinations
General objective	To demonstrate the efficacy of the novel therapy	To compare the effectiveness of available therapies
Eligibility criteria	Strict, more homogeneous	Less strict, closer to 'real life'
Timing of randomization	As close to interventions as possible	Following practical constraints
Intensity of follow-up and patient assessment	More intense	Less intense
Data collection	More intense	Less intense
Analysis principles	Some restrictions may be applied to the intent-to-treat population	Intent-to-treat population

such multi-arm trials may require special attention (e.g., see [8] for when to adjust for multiplicity, as well as the HERA trial for an example of such adjustment [9]). Even though multi-arm trials enroll more patients than two-arm trials, it has been argued that overall resources could be saved by having different regimens (and hence perhaps different sponsors) compared with a common control arm in multi-arm trials [8]. With further statistical refinements and planned adaptations, so-called multi-arm, multi-stage trials may allow simultaneous assessment of various novel treatments against a single control arm, with discontinuation of arms that do not show sufficient promise [10]. This design has been used, for example, in the STAMPEDE trial in prostate cancer [11].

12.2.3 Choice of the Control Arm

Ideally, the control arm should be the best available standard of care at the time of trial design and conduct. However, improvements in cancer care may lead to difficulties in choosing control arms for phase III trials, especially if they are expected to last for many years. At times, the best control may be a “moving target”, and protocol amendments may be required during the study to accommodate for changes in the standard of care after the trial was started. In oncology, most phase III trials have active comparators, with placebo or observation alone being used less frequently. More recently, treatment of physician’s choice (as done, e.g., in the EMBRACE trial [12]) has emerged as a means to ensure trial feasibility in settings for which multiple active comparators are available.

12.2.4 Sample Size

For ethical, scientific, and monetary reasons, sample-size estimation is a key aspect of any clinical trial [13, 14]. However, determining the sample size assumes special importance in phase III trials, given their attempt to allow for sound statistical conclusions upon their completion. These conclusions—whether positive or negative—are only warranted if the sample size was estimated correctly. I believe the elements required for sample-size estimation may be conveniently embodied in what may be called the “ABCDE of sample-size calculation”. Table 12.2 describes the components of the ABCDE (alpha, or the type-I error; beta, or the type-II error; control, or the expected result in the control arm; dropout, or the expected rate of loss to follow-up or loss of data; experimental, or the desired result in the experimental arm, using the same parameter type as for C), most of which are well-known to physicians and allow them to communicate properly with statisticians responsible for the calculations. Additional elements may need to be taken into account in some settings; e.g., the accrual rate (or duration) and the duration of follow-up are required to estimate sample sizes for time-to-event endpoints. Moreover, group-sequential trials, several adaptive designs, biomarker-based trials, factorial trials, and

Table 12.2 The ABCDE of sample-size calculation

Component	Explanation	Comments
A	Alpha, or the type-I error	The chance of a false-positive result is conventionally set at 5% (two-sided) for superiority phase III trials; for non-inferiority trials, one-sided alpha of 2.5% or 5% is commonly used
B	Beta, or the type-II error	The chance of a false-negative result is more often set at 10–20%, thus leading to power of 80–90%
C	Control, or the expected result in the control arm	Ideally, this should come from the published literature or from data from the same institution(s) as the one(s) conducting the trial; this may be a mean (with standard deviation), a proportion, a median, a hazard rate, or other parameters depending on the setting and type of primary endpoint variable
D	Dropout, or the expected rate of loss to follow-up or loss of data	This is commonly set at 10%, but may depend on the clinical setting
E	Experimental, or the desired result in the experimental arm, using the same parameter type as for C	This educated guess should be based on plausible—and preferably clinically relevant—treatment effects that can be expected; very often, it comes from historical practice, but pilot data from the institution(s) may also help

cluster-randomized trials require considerations that go beyond the elements of the ABCDE. It should be noted that the exercise of estimating the required sample size yields the number of evaluable patients (or events, in the case of time-to-event endpoints), which may be different from the total number of patients to be enrolled.

Several review articles on sample-size calculation for clinical trials are available, some of them presenting tables with numbers of required patients for selected ABCDE parameters [13–19]. The most critical aspects of this exercise are the choice of the primary endpoint and the determination of what difference between the control and experimental arm the trial will target. This difference is the "treatment effect", which, in cancer phase III trials, often corresponds to the hazard ratio. Other parameters being equal, the smaller the treatment effect sought, the larger the sample size required. This sounds counterintuitive at first, but obviously it does not take many patients to find a statistically significant difference when this is large. The expected result for the primary endpoint in the control arm is often based on the literature and corresponds to the statistical notion of the null hypothesis. The desired result in the experimental arm, which corresponds to the statistical notion of the alternative hypothesis, is in fact an educated guess that should be plausible and clinically relevant, although very often it is based on historical practice in the same setting as the trial. Of note, most superiority trials do not test for the superiority of the experimental arm, but rather seek differences in both directions (i.e., superiority and inferiority). In other words, superiority trials use two-sided statistical tests, and superiority is claimed when the results favor the experimental arm with statistical significance.

12.2.5 Mechanics of Randomization

Randomization aims at balancing treatment arms for known and, more importantly, unknown prognostic factors, thus avoiding selection bias. It is worth mentioning that a proper randomized trial must display two essential features: the unpredictability of the allocation sequence and proper concealment of such sequence when it is determined before patient enrollment [20]. Although there is some disagreement among authors, simple randomization is generally recommended only for very large trials. In oncology, where trials are typically not very large and often multicenter, randomization is more often implemented using two additional features: blocking and stratification. Blocked randomization consists in generating randomization sequences that ensure the intended balance in *patient numbers* after the enrollment of a fixed number of individuals at any point in time during accrual. This is important for the trial overall and for centers enrolling small numbers of patients. In unblinded trials, the block size should not be known to investigators, in order to avoid predictability of the last few patients in each sequence. Stratification, which aims at ensuring balance in known *prognostic factors* at any point in time during accrual, should be considered in every trial. However, only a few factors can be used to stratify patients, otherwise the randomization scheme becomes too complex and statistically inefficient. In oncology, the study center is often used as a stratification factor to control for ancillary patient care unrelated to the trial, so that in most cases only one or two additional factors may be considered. An alternative to stratification is minimization, a dynamic randomization method for which there is no predefined allocation sequence, and each patient is randomized in a way that minimizes the imbalance in predefined prognostic features across treatment arms. Another form of dynamic randomization, also called outcome-adaptive randomization, takes account of ongoing results in a trial in order to allocate patients to the arm(s) that are performing better; although advocated by some, this method is very controversial and should not be used routinely [21].

Although randomization is more often done following a 1:1 ratio, unequal randomization is very useful in settings for which there is already substantial information about the control arm, and thus having more patients (e.g., twice as many) in the experimental arm may bring efficiency at the expense of a small increase in sample size (of the order of 10–15% for 2:1 ratios). Several pivotal superiority phase III trials in oncology have used a 2:1 randomization (see [12, 22] for examples). A final word about randomization relates to settings for which it is not possible to randomize within institutions or departments (called clusters for this purpose) that have a preference for one of the competing interventions in a phase III trial. In this case, the units of randomization are no longer the patients, but the clusters, such that all patients in a cluster receive the same intervention. Owing to correlation issues within clusters, cluster-randomized trials require special precautions both for design and for analysis, but this design is almost never used in oncology.

One special type of randomization applies to so-called factorial trials, in which the primary aim is to assess two factors simultaneously (although more than two may be tested, in oncology only factorial trials testing two factors have been used, to my knowledge). In these 2 × 2 trials, patients are randomized twice: once between the two interventions in one factor of interest, and then to the two interventions in the second factor [23]. A typical example is the REAL-2 trial, in which patients with advanced esophagogastric cancer treated with epirubicin (E) were randomized once to cisplatin (C) versus oxaliplatin (O) (the "platinum factor") and once to fluorouracil (F) versus capecitabine (X) (the "fluoropyrimidine factor") [24]. The goal was to make comparisons within each factor separately, and not across the four arms thus formed (ECF, EOF, ECX, EOX), as explained below.

12.2.6 Blinding and Other Assessment Issues

Whenever feasible, blinding is a desirable feature that aims at removing subjectivity in the assessment of outcomes, especially for endpoints that are less objective, such as evaluation of toxicity, thus avoiding observation bias. Moreover, blinding may ensure that ancillary care or any other action that may affect outcomes is applied by investigators with no preference for one of the treatment arms [7]. Although most phase III trials in oncology are not double-blinded from the point of view of both patients and physicians, blinded assessment of responses and progression (by independent reviewers or radiologists) is often implemented. Regardless of the use of blinding, it is imperative that the same schedule of assessment is used in all arms of a phase III trial. Both for safety and efficacy endpoints, differential assessment across arms may introduce bias that can seriously threaten the conclusions of the study.

12.3 Endpoints for Phase III Trials

12.3.1 Function and Hierarchy of Endpoints

Endpoints are the metrics chosen to represent the outcome variables of interest and whose differential change after treatment allows for comparisons between treatment arms. Mostly because of statistical concerns with false-positive results, there needs to be a hierarchy of endpoints within a given trial. This hierarchy reflects the perceived importance of the endpoints, historical practice in the field, and regulatory constraints [25]. The primary endpoint serves two very important functions: to allow estimation of the sample size and to ascertain results as positive or negative upon study completion. From the regulatory and statistical points of view, secondary endpoints have an exploratory role; moreover, they should be limited in number and should be seen as supportive evidence regarding the primary endpoint [26].

12.3.2 Efficacy Endpoints

Although three general types of variables are used as efficacy endpoints in clinical trials—numerical, categorical, and time-to-event—only the latter two are used frequently in oncology. Categorical variables include response rates, the clinical benefit rate, and survival rates at specified landmarks (e.g., 1-year survival rate), whereas all "survival" endpoints are of the time-to-event type. The objective response rate is defined as the proportion of patients achieving confirmed complete or partial responses as assessed by valid imaging methods [27], but response rates may also be defined on the basis of tumor markers (e.g., prostate-specific antigen [28] or CA-125 [29]), pathological assessment [30], or for specific settings (e.g., neuro-oncology [31]). The clinical benefit rate has no uniform definition, with some authors considering it as the proportion of patients with complete or partial responses or with stable disease of any duration, whereas others specify a minimum duration (usually of ~6 months) of disease stability. The definitions of the most common survival endpoints are displayed in Table 12.3. For such definitions, the events of interest and the censoring rules are pre-specified and lead to sometimes subtle differences between different endpoints. Moreover, efforts are sometimes required to ensure standardization of such definitions (see, e.g., [32] and other articles in the series).

There has been a heated debate in the literature about the relative merits of overall survival (OS) or progression-free survival (PFS) as the most adequate primary endpoint in the metastatic setting. Both of these endpoints have pros and cons, which are essentially opposed to each other: OS is objective and clearly relevant to patients, but is inefficient to statistically and prone to "contamination" by post-trial therapy, whereas PFS is prone to measurement error and is of doubtful relevance to patients,

Table 12.3 Definitions of selected time-to-event (survival) endpoints in oncology

Endpoint	Event(s) of interest	Reasons for censoring	Setting
Overall survival	Death from any cause	End of follow-up (i.e., patient still alive) or loss to follow-up	C and P
Cancer-specific survival	Death from cancer	End of follow-up (i.e., patient still alive), death from other causes, or loss to follow-up	C and P
Progression-free survival	Disease progression or death from any cause	End of follow-up (i.e., patient still alive and without progression) or loss to follow-up	P
Time to tumor progression	Disease progression	End of follow-up (i.e., patient without progression), death without previous documentation of disease progression, or loss to follow-up	P
Disease-free survival	Disease recurrence or death from any cause	End of follow-up (i.e., patient still alive and without recurrence) or loss to follow-up[a]	C
Time to treatment failure	Disease progression, treatment toxicity, patient preference, or death from any cause	End of follow-up (i.e., with no event of interest) or loss to follow-up	P
Duration of response	Disease progression (from the date of response documentation, only for responders)	End of follow-up (i.e., patient with no disease progression), death without prior documentation of disease progression, or loss to follow-up	P

C curative (adjuvant or neoadjuvant), *P* palliative (advanced or metastatic)

[a] Several definitions are available for variations of disease-free survival according to the setting

but is more statistically efficient and is not subject to the effect of post-trial therapy [33]. Although there is still no universal consensus about this issue, PFS is currently the most frequent primary endpoint used in various settings, especially in the first-line setting and when salvage therapies are available. However, the frequency of use of different efficacy endpoints varies across tumor types and over time.

12.3.3 Other Endpoints

Safety endpoints in oncology frequently equate with the rates of adverse events—as assessed by the Common Terminology Criteria for Adverse Events [34]—and the rates of laboratory abnormalities. International Conference of Harmonization guidelines should also be followed with regard to serious adverse events and other pertinent issues. With very few exceptions (see, e.g., [35]), safety is almost always a secondary endpoint in phase III trials, albeit a key one. Likewise, quality-of-life endpoints are seldom used as primary variables in phase III trials (see, e.g., [36] for an exception). Other types of endpoints are used on occasion, including cost-effectiveness [37] and the benefit rate defined on the basis of symptom control [38].

12.4 Selected Issues in Analysis and Interpretation

12.4.1 Analysis Populations

In a phase III trial, at least three populations are commonly defined and may be used for various analyses: the intent-to-treat (ITT), safety, and per-protocol populations. However, there should always be one primary analysis, which allows for conclusions about whether the study is positive or negative. As a general rule, the primary efficacy analysis of a superiority phase III trial should be made using the ITT population, defined as all randomized patients grouped according to their allocation. Therefore, patients who are not treated or are excluded from the study for any reason, and patients treated in the wrong arm (see Fig. 1 in [39] for proof that this is not such a rare occurrence) are analyzed in the arm to which they were randomized as if they had been treated. For example, in a phase III superiority trial whose primary endpoint is response rate, a patient who was randomized but died before response evaluation is considered to be a non-responder. This aims at eliminating selection bias introduced by the exclusion of patients that differ systematically from the non-excluded patients, often in unknown ways. A key recommendation in the analysis of randomized trials is that no patients should be excluded after they have been randomized, as these exclusions undermine the validity of results [6]. Losses to follow-up may be dealt with through censoring, but every effort should be made to minimize them. The safety analyses are usually conducted in the safety population, defined as all patients in the ITT population who received at least one dose of the study drug, and grouped according to what they in fact received. Finally, per-protocol populations may be defined more strictly (e.g., as patients in the ITT population who received a minimum number of cycles and underwent a minimum number of assessments), but the analyses based on these populations are prone to bias, and such analyses are therefore exploratory in a superiority trial.

Special precautions are needed for the analysis of factorial trials. The higher statistical efficiency of these trials stems from the fact that the primary comparison is within each factor separately. This leads to much smaller sample sizes than if the same trial was designed with four arms in the attempt to make cross-arm comparisons or even only comparisons between each experimental arm and the control arm (as done, e.g., in the ECOG [Eastern Cooperative Oncology Group] trial 1594 [40]). The price to pay, however, is that comparison between arms is exploratory and typically underpowered. A second caveat in these trials is that interactions between interventions may preclude the primary analysis, because these trials are designed under the additivity—i.e., no interaction—assumption [23]. In the ECOG 1199 trial, for example, the goal was to compare schedules (weekly versus every 3 weeks) and taxanes (paclitaxel versus docetaxel), but an interaction between the schedule and the taxane compromised the primary analysis [41].

12.4.2 Subgroup Analyses

Once the overall efficacy results of a phase III trial have been assessed, it is difficult to resist the temptation to conduct assessments of the treatment effect for specific endpoints in subgroups of patients defined by baseline characteristics [42]. These subgroup analyses suffer from two inherent problems: first, they are prone to an increased type-I error, given the multiplicity of analyses conducted; second, they have less power than the study overall, thus being prone to inflation of the type-II error. Despite these caveats, subgroup analyses may be informative as a means to generate hypotheses for further testing. At times, they may even uncover important treatment-by-covariate interactions that lead to changes in patient management, as was the case for patients with KRAS mutations in metastatic colorectal cancer treated with anti-epidermal growth factor receptor antibodies [43]. More often, however, subgroup analyses suggest issues that prove to be irrelevant, such as progesterone positivity for the efficacy of aromatase inhibitors [44]. Key aspects in the interpretation of reported subgroup analyses are the availability of information about how many such analyses were planned and how many were conducted; the focus on tests for interaction rather than *P* values within subgroups; the consistency of results if different endpoints were analyzed; the consistency of results in similar studies; and their biological plausibility.

12.4.3 Interim Analyses for Efficacy

Formal interim analyses for efficacy in group-sequential phase III trials should be distinguished from other types of interim looks at data from ongoing trials [45]. The former are not mandatory, but are an important safeguard for patients and require statistical rigor; the latter are probably frequent, but they can yield misleading results and should be avoided. Formal interim analysis, conducted in the context of an independent data monitoring committee (also called a data and safety monitoring board), aims at stopping patient enrolment or treatment if there is sufficient evidence of difference between arms, or if continuation of the trial would be futile because the chance of finding such difference at the end is too small given the available data. Such analyis is done following a pre-specified statistical plan that takes account of multiplicity and thus requires very low *P* values for decisions about efficacy and additional considerations for futility. Several trials have been stopped early for efficacy, and examples include the HERA trial [9] and the PREVAIL trial [46]; stopping for futility is less frequent, but examples can be found [47, 48].

12.4.4 Beyond Statistics

Statisticians are practically unanimous in affirming that the final interpretation of any clinical trial rests on scientific judgment more than on statistical criteria [49]. Nevertheless, the medical community seems to remain convinced that a significant *P* value represents the ultimate proof of success of a superiority phase III trial. This view is incorrect for at least three reasons. The first is that a successful clinical trial is not one that yields a positive result, but rather one that answers the main scientific question posed by the trial, whether this answer is positive or negative. Secondly, any statistical conclusion is subject to error, because samples rather than populations are studied. Finally, the dichotomy between "positive" and "negative" trials is artificial, and several considerations are in order in either of these two situations [50, 51]. Many of these considerations are of a medical nature and relate to eligibility criteria, treatment plan, choice of primary and secondary endpoints, toxicity, etc. From a statistical point of view, it can be added that the magnitude of the difference (the "size" of the treatment effect) and the confidence interval for the treatment effect, rather than *P* values, are clearly more informative, respectively, as a measure of benefit and as the basis for inference about the results.

12.5 Non-inferiority Trials

Non-inferiority trials are usually designed to compare two interventions with similar mechanisms of action, but possible differences in patient convenience. A typical example is the comparison between an oral and an intravenous agent from the same class [52]. In these cases, it is essential to assess whether the new treatment is not unacceptably worse than a treatment already in use [53]. In order to demonstrate that a new treatment is clinically equivalent to an existing standard, it is inappropriate to declare equivalence on the basis of a non-significant *P* value from a superiority comparison. Rather, non-inferiority has to be demonstrated formally by showing that the new treatment is not worse than the standard by more than a specified margin, which is called the non-inferiority margin. Figure 12.1 illustrates the results

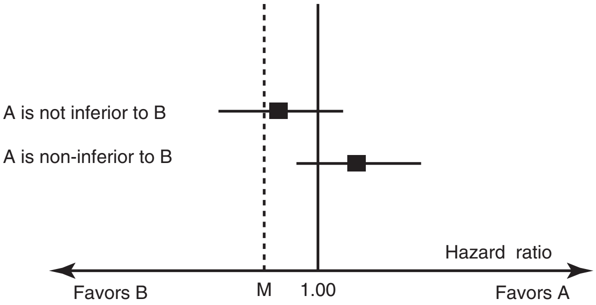


Fig. 12.1 Graphical display of results from two fictitious trials, both comparing the new treatment A and the old treatment B; in the first case, results represent a negative superiority trial (A is not superior but is also *not inferior* to B), because the 95% confidence interval (the horizontal line) for the hazard ratio (the square) crosses the line of unity; in the second case, a positive non-inferiority trial shows that A is *non-inferior* to B, because the lower limit of the 95% confidence interval for the hazard ratio is above M, the margin of non-inferiority (which obviously would only be used in the non-inferiority trial)

from two fictitious trials, both comparing treatments A and B; in the first case, A is *not inferior* to B, but in the second, A is *non-inferior* to B. What appears to be a semantic game in fact discloses the differences between a negative superiority trial and a positive non-inferiority trial.

Determining this non-inferiority margin is the greatest challenge in the design, conduct, and interpretation of non-inferiority trials. The ideal non-inferiority trial would have a placebo control, in order to ensure that the new treatment is not only non-inferior to the old, but also superior to placebo. In oncology, this is usually inappropriate, so the choice of the non-inferiority margin is often made on a historical rather than a statistical basis [54, 55]. For non-inferiority trials, a one-sided type-I error rate of 2.5% is often recommended, since it corresponds to a two-sided error of 5%. For the interpretation of results, the lower limit of the 95% confidence interval for the difference between arms should be above the non-inferiority margin [56]. Unlike superiority trials, there is some controversy about whether the ITT or the per-protocol population should be used for the primary analysis of a non-inferiority trial. In a non-inferiority trial, the chance of finding similar effects for two treatments is potentially increased by the use of the per-protocol population, and therefore some authors have argued that the use of this population is a more conservative and preferred approach in this setting [57]. However, there is no consensus about this issue, and the vast majority of non-inferiority cancer trials give precedence to the ITT population for their primary analyses [54]. Ideally, non-inferiority cancer trials should report both ITT and per-protocol analyses.

Conclusion

Phase III trials are the ideal scenario for comparing two or more competing therapies. Ideally, phase III trials should have two arms, use the best available standard of care as the control arm, and have their sample sizes determined accurately to allow for sound statistical conclusions. Deviations from these ideal features are useful, but require a good rationale and sound ethical and statistical bases. Blocked randomization, stratification, minimization, blinding, and analyses based on the intent-to-treat principle are additional safeguards to allow for unbiased conclusions from phase III trials. Non-inferiority and factorial trials present some additional features that warrant special attention in their design and interpretation.

References

1. Buyse ME, Staquet MJ, Sylvester RJ. Cancer clinical trials: methods and practice. New York: Oxford University Press; 1984.
2. Girling D, Parmar M, Stenning S, et al. Clinical trials in cancer. Oxford: Oxford University Press; 2003.
3. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20:637–48.
4. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009;180:E47–57.
5. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med. 2016;375:2293–7.
6. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer. 1976;34:585–612.
7. Green SB. Randomized clinical trials: design and analysis. Semin Oncol. 1981;8:417–23.
8. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. Clin Cancer Res. 2008;14:4368–71.
9. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353:1659–72.
10. Sydes MR, Parmar MK, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. Trials. 2009;10:39.
11. James ND, Sydes MR, Clarke NW, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. Lancet. 2016;387:1163–77.
12. Cortes J, O'Shaughnessy J, Loesch D, et al. Eribulin monotherapy versus treatment of physician's choice in patients with metastatic breast cancer (EMBRACE): a phase 3 open-label randomised study. Lancet. 2011;377:914–23.
13. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365:1348–53.
14. Curran D, Sylvester RJ, Hoctin Boes G. Sample size estimation in phase III cancer clinical trials. Eur J Surg Oncol. 1999;25:244–50.
15. Fayers PM, Machin D. Sample size: how many patients are necessary? Br J Cancer. 1995;72:1–9.
16. Florey CD. Sample size for beginners. BMJ. 1993;306:1181–4.
17. Rohrig B, du Prel JB, Wachtlin D, et al. Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107:552–6.
18. Whitley E, Ball J. Statistics review 4: sample size calculations. Crit Care. 2002;6:335–41.
19. Simon R. Size of phase III cancer clinical trials. Cancer Treat Rep. 1985;69:1087–93.
20. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359:614–8.
21. Korn EL, Freidlin B. Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. J Natl Cancer Inst. 2017;109. https://doi.org/10.1093/jnci/djx013.

22. Shepherd FA, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med. 2005;353:123–32.
23. Green S, Liu PY, O'Sullivan J. Factorial design considerations. J Clin Oncol. 2002;20:3424–30.
24. Cunningham D, Starling N, Rao S, et al. Capecitabine and oxaliplatin for advanced esophagogastric cancer. N Engl J Med. 2008;358:36–46.
25. Sargent D. General and statistical hierarchy of appropriate biologic endpoints. Oncology (Williston Park). 2006;20:5–9.
26. European Medicines Agency. ICH Topic E9 - Statistical Principles for Clinical Trials. Note for Guidance on Statistical Principles for Clinical Trials (CPMP/ICH/363/96). Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf. Accessed 21 Apr 2017.
27. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–47.
28. Scher HI, Halabi S, Tannock I, et al. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the Prostate Cancer Clinical Trials Working Group. J Clin Oncol. 2008;26:1148–59.
29. Rustin GJ, Vergote I, Eisenhauer E, et al. Definitions for response and progression in ovarian cancer clinical trials incorporating RECIST 1.1 and CA 125 agreed by the Gynecological Cancer Intergroup (GCIG). Int J Gynecol Cancer. 2011;21:419–23.
30. O'Connell MJ, Colangelo LH, Beart RW, et al. Capecitabine and oxaliplatin in the preoperative multimodality treatment of rectal cancer: surgical end points from National Surgical Adjuvant Breast and Bowel Project trial R-04. J Clin Oncol. 2014;32:1927–34.
31. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in Neuro-Oncology Working Group. J Clin Oncol. 2010;28:1963–72.
32. Gourgou-Bourgade S, Cameron D, Poortmans P, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). Ann Oncol. 2015;26:873–9.
33. Saad ED, Buyse M. Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents. Ann Oncol. 2016;27:373–8.
34. U.S. Department of Health and Human Services. National Cancer Institute. Cancer Therapy Evaluation Program. Common Terminology Criteria for Adverse Events (CTCAE). Version 4.0. Available at http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf. Accessed 21 Apr 2017.
35. Stathopoulos GP, Antoniou D, Dimitroulis J, et al. Liposomal cisplatin combined with paclitaxel versus cisplatin and paclitaxel in non-small-cell lung cancer: a randomized phase III multicenter trial. Ann Oncol. 2010;21:2227–32.
36. Gronberg BH, Bremnes RM, Flotten O, et al. Phase III study by the Norwegian Lung Cancer Study Group: pemetrexed plus carboplatin compared with gemcitabine plus carboplatin as first-line chemotherapy in advanced non-small-cell lung cancer. J Clin Oncol. 2009;27:3217–24.
37. Vergnenegre A, Corre R, Berard H, et al. Cost-effectiveness of second-line chemotherapy for non-small cell lung cancer: an economic, randomized, prospective, multicenter phase III trial comparing docetaxel and pemetrexed: the GFPC 05-06 study. J Thorac Oncol. 2011;6:161–8.
38. Burris HA III, Moore MJ, Andersen J, et al. Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. J Clin Oncol. 1997;15:2403–13.
39. Geyer CE, Forster J, Lindquist D, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med. 2006;355:2733–43.
40. Schiller JH, Harrington D, Belani CP, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. N Engl J Med. 2002;346:92–8.
41. Sparano JA, Wang M, Martino S, et al. Weekly paclitaxel in the adjuvant treatment of breast cancer. N Engl J Med. 2008;358:1663–71.
42. Sun X, Ioannidis JP, Agoritsas T, et al. How to use a subgroup analysis: users' guide to the medical literature. JAMA. 2014;311:405–11.

43. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol. 2008;26:1626–34.
44. Dowsett M, Cuzick J, Wale C, et al. Retrospective analysis of time to recurrence in the ATAC trial according to hormone receptor status: an hypothesis-generating study. J Clin Oncol. 2005;23:7512–7.
45. Whitehead J. Interim analyses and stopping rules in cancer clinical trials. Br J Cancer. 1993;68:1179–85.
46. Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. N Engl J Med. 2014;371:424–33.
47. Blay JY, Shen L, Kang YK, et al. Nilotinib versus imatinib as first-line therapy for patients with unresectable or metastatic gastrointestinal stromal tumours (ENESTg1): a randomised phase 3 trial. Lancet Oncol. 2015;16:550–60.
48. Rugo HS, Barry WT, Moreno-Aspitia A, et al. Randomized phase III trial of paclitaxel once per week compared with nanoparticle albumin-bound Nab-paclitaxel once per week or ixabepilone with bevacizumab as first-line chemotherapy for locally recurrent or metastatic breast cancer: CALGB 40502/NCCTG N063H (Alliance). J Clin Oncol. 2015;33:2361–9.
49. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.
50. Pocock SJ, Stone GW. The primary outcome is positive - is that good enough? N Engl J Med. 2016;375:971–9.
51. Pocock SJ, Stone GW. The primary outcome fails - what next? N Engl J Med. 2016;375:861–70.
52. Twelves C, Wong A, Nowacki MP, et al. Capecitabine as adjuvant treatment for stage III colon cancer. N Engl J Med. 2005;352:2696–704.
53. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials. 2011;12:106.
54. Saad ED, Buyse M. Non-inferiority trials in breast and non-small cell lung cancer: choice of non-inferiority margins and other statistical aspects. Acta Oncol. 2012;51:890–6.
55. Riechelmann RP, Alex A, Cruz L, et al. Non-inferiority cancer clinical trials: scope and purposes underlying their design. Ann Oncol. 2013;24:1942–7.
56. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J Clin Oncol. 2008;26:3543–51.
57. Piaggio G, Elbourne DR, Altman DG, et al. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006;295:1152–60.

Phase IV Trials: Interventional and Non-interventional Studies

13

Deise Uema, Cheng Tzu Yen, Axel Hinke,
and Gilberto de Castro Jr.

13.1 Introduction

This chapter focuses on phase IV trials, which are studies designed in the post-marketing scenario to certify approved medications in the real-world population [1, 2]. Phase IV refers to large studies, interventional or non-interventional, that are often used to assess serious adverse effects in a sizeable population, but that are sometimes also used to approve additional uses of a drug or to introduce physicians and patients to new treatments [1]. While only 20% of the drugs that enter phase I trials are approved for marketing, approximately 20% of new medications acquire new black box warnings after commercialization, and around 4% of drugs are withdrawn for safety reasons [3, 4]. Here we will discuss the main phase IV study designs, and the potential advantages and limitations of these methodologies.

13.2 Definitions of Phase IV Studies

A phase IV study is a clinical study where the investigational therapy includes the use of a licensed drug or device, as seen in Fig. 13.1 [1, 2]. According to the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA), the definitions of phase IV studies are as follows:

D. Uema, M.D. • G. de Castro Jr., M.D., Ph.D. (✉)
ICESP—Medicine School of University of São Paulo, São Paulo, SP, Brazil

Sirio Libanês Hospital, São Paulo, SP, Brazil

C. T. Yen, M.D.
Oswaldo Cruz German Hospital, São Paulo, SP, Brazil

A. Hinke, Ph.D.
WiSP Research Institute, Langenfeld, Germany

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_13

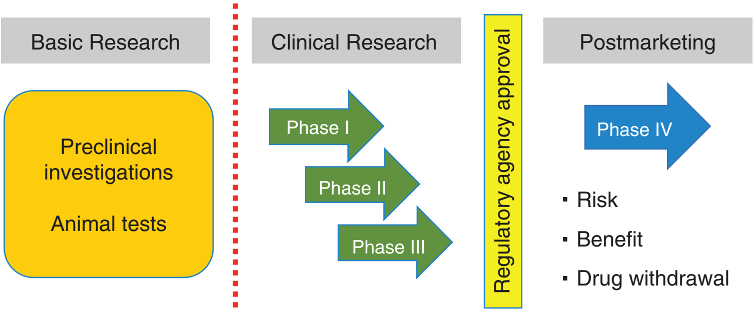


Fig. 13.1 Overview of drug development

- FDA: Phase IV studies are post-marketing studies that are imposed upon a pharmaceutical firm as a condition for drug approval. Phase IV trials are carried out once the drug or device has been approved by the FDA during the Post-Market Safety Monitoring (Clinical Research–fda.gov, https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm—accessed April 16th, 2017) [5].
- EMA: Studies in phase IV are all the studies (other than routine surveillance) that are performed after drug approval and related to the approved indication. They are studies that were not considered necessary for approval but are often important for optimizing the drug's use. They may be of any type, but should have valid scientific objectives. Commonly conducted studies include additional drug-drug interaction, dose-response, or safety studies and studies designed to support use according to the approved indication, e.g., mortality/morbidity studies and epidemiological studies (http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002877.pdf—accessed April 16th, 2017) [6].

Since phase II and phase III trials use planned but limited samples from the target population in a limited timeframe, some events or interactions may not present before exposure to the real world and findings may lack adequate external validity. These studies aim to demonstrate efficacy, and to this end, they seek internal validity using carefully selected individuals under ideal circumstances [2]. Internal validity, defined as the reliability of the observed outcomes and proper control of bias [1, 7], is usually obtained using selective inclusion criteria, therefore creating a strict study population. Conversely, the concept of external validity relies on the notion of generalizability; that is, to what extent the results obtained in clinical trials may be extrapolated to a broader selection of patients and settings [1, 7].

In this sense, phase IV studies are usually conducted in order to optimize the use of treatments (dosage, duration of therapy, implementation in complex drug strategies), to adapt indications including groups of individuals that were not represented in pre-approval studies, to explore potential effect modifiers (patient characteristics or comorbidities that may influence pharmacokinetics), or to observe safety issues that phase III trials were not powered to evaluate (routine, rare, or delayed side effects; interactions with other therapies; addiction and abuse) [1]. Phase IV studies can provide natural intriguing insights about unknown interactions involving populations that are heterogeneous (e.g., in regard to genetics, habits, and comorbidities) and thousands of medical products widely used in real-life.

In addition to the evaluation of the biological effects of new therapies, phase IV trials can be designed to assess cost-effectiveness or marketing proposals (acceptance and compliance by physicians and patients) [1]. Phase IV studies are particularly useful in oncology because they can provide effective treatments to patients with refractory metastatic cancers for which there are no available therapeutic options.

In the United States, Expanded Access Programs (EAPs), also known as "Compassionate Use", are study models that allow patients with serious diseases or conditions; for example, cancer patients, to have access to new drugs, biologics, or medical devices [8]. EAPs are considered when patient enrollment in a clinical trial is not feasible (e.g., the patient is not eligible or there are no ongoing trials) [8] and the intention of these models is to treat patients instead of obtaining data on efficacy and safety [9, 10]. Despite the greater flexibility in inclusion criteria in EAPs, the FDA also demands the sponsor (or the clinical investigator) to demonstrate that there are no equivalent available treatments and to provide a clinical protocol, approved by institutional review boards and subject to informed consent [8].

Differently from phase III trials, which are mandatory for the approval of a new drug or therapy, only 25% of marketed drugs move forward to phase IV studies [11]. Recently, classical phase IV trials have been requested by regulatory agencies such as the FDA and EMA, but these trials are still sparsely regulated [11, 12]. Phase IV trials also do not have a definite standard approach, preferred study design, or particular statistical method.

13.3 Designs of Phase IV Studies

Although randomized clinical trials are the default study design in phase III trials, post-marketing trials can use several different approaches, each with its particular methodology and limitations. The FDA has mandated or negotiated studies ranging from controlled trials to observational studies, drug-drug interaction studies, or special population studies [1, 12, 13]. Table 13.1 exemplifies the most common designs of phase IV trials currently used.

Table 13.1 Most common phase IV (interventional and non-interventional) study designs [12, 13]

Design	Objectives	Examples
Non-interventional study	Conducted to assess a treatment's safety, tolerability, and effectiveness in real-world conditions. There are no additional tests or visits other than those that would be done in usual practice. Informed consent primarily encompasses a privacy clause. Despite being a simple trial model, the data collection can be flawed, compromising the evaluation of outcomes (such as survival or adverse events)	Ofatumumab was evaluated in a retrospective, phase IV non-interventional, observational study in heavily pre-treated patients with poor-prognosis chronic lymphocytic leukemia. Data from 103 patients treated with ofatumumab outside phase II or phase III ofatumumab-based trials were collected and analyzed for progression-free survival and overall survival. Owing to the low reporting of adverse and infusion reactions in patients' records, the toxicity profile assessment may have been compromised [14]
Large simple trial	Such trials combine aspects of a randomized clinical trial and an observational study. A large and more heterogeneous number of patients is accrued to the study, aiming to maximize the external validity and generalizability of the therapy investigated. Differently from a non-interventional study, this type of trial uses simplified data entry and management, reducing the costs compared with those of complex phase III trials. This type of trial is the only one categorized as a "phase IV trial" in the European Union	The SAiL (Safety of Avastin in Lung) study is a good example of a large simple trial in oncology. This was an open-label, multicenter, phase IV study conducted with the purpose of evaluating the safety and efficacy of first-line bevacizumab combined with standard chemotherapy in a large population with advanced non-squamous non-small-cell lung cancer. There were 2212 individuals enrolled, including elderly patients and patients with a performance status of 2, characteristics that are common in clinical practice populations, but not in phase III clinical trials. The final data confirmed a manageable safety profile and efficacy in a more comprehensive cohort of patients than that in a phase III trial [15]
Post-marketing surveillance	Post-marketing surveillance helps to detect rare adverse reactions, especially those with a frequency of less than 1 in 3000–5000, that are unlikely to appear in phase I-III trials. Besides safety issues, post-marketing surveillance can test the tolerability and effectiveness of an intervention in the real world; that is, the effect of the intervention out of the controlled environment of clinical trials	A study in 470 patients with imatinib-resistant/-intolerant gastrointestinal stromal tumor, conducted to expand the sunitinib safety database. No new side effects were reported, but a higher rate of adverse events ≥3 was observed in the early exposure to the drug, different from what was described in phase III trials [16]

Adverse event monitoring	Some rare adverse events are unlikely to occur before 30,000–65,000 individuals are exposed to a new intervention. Considering this, safety monitoring should continue for the lifetime of a drug. The reporting of serious and/or unexpected adverse reactions can help regulatory authorities to evaluate for possible causal relationships and even consider removal from the market in some situations. This monitoring is usually done by continuous pharmacovigilance on the part of industries, but can also emerge from trials or vigilance systems. In the United States, the FDA has implemented “MedWatch”, a voluntary system for the reporting of adverse events, where patients, providers, and manufacturers can include their experiences. The purpose of this system is to facilitate the identification of risk signals. One of the main problems with MedWatch is the high underreporting rate, which can compromise the evaluation of risks and delay possible safety actions	Vismodegib, a medication for basal cell carcinomas, had uncommon liver toxicity, evaluated by the MD Anderson Cancer Center. The investigators evaluated the current literature and were able to find only two reports of severe liver dysfunction with the use of vismodegib. However, after searching the FDA Adverse Event Reporting System, they found 94 reports of adverse events, 35 of which were serious. Previous trials included patients with normal liver function, a factor that could have hampered the evaluation of hepatic toxicity. This study was able to identify the safety profile of vismodegib, changing the use of this drug in patients with some degree of hepatic impairment [17]
Retrospective case-control studies	Case-control studies are used to assess possible rare side effects. They can help to determine the likelihood of the association of a drug with an adverse event	A case-control study, conducted by Patel et al., evaluated the incidence of intestinal intussusception after the first dose of rotavirus vaccine. Cases of intussusception were identified, independently of the patients’ vaccination status, and compared in up to four controls matched by dates of birth and neighborhood. The authors found an association between the first dose of rotavirus vaccination and intussusception among infants in a Mexican population (which was corroborated by findings of the manufacturer in two different populations), but not in Brazilian children. The authors hypothesized that this difference might be attributable to the concomitant administration of oral poliovirus vaccine in Brazil, which reduces the replication of rotavirus strains in the intestines, thereby reducing the inflammatory response in intestinal lymphatic tissue and the risk of intestinal obstruction [18]

(continued)

Table 13.1 (continued)

Design	Objectives	Examples
Drug utilization studies	Drug utilization studies describe how a drug is marketed and actually prescribed within a population and how these factors can influence outcomes	Duran et al. performed a systematic review of drug utilization studies in Latin America. The authors noted a paucity of available data on drug consumption, especially in the public healthcare sector; such paucity can impair the validity of conclusions, considering the extensive extrapolation of data [19]
Registry studies	Registry studies are prospective observational studies of patients with similar risks or diseases that can present a precise overall picture of clinical practice, patients, and outcomes	Driessen et al. reported short- and long-term efficacy and safety data on biologics used for psoriasis, especially etanercept. The authors prospectively collected registry data from 118 patients in a single-center outpatient clinic. Short-term and long-term etanercept efficacy analysis showed an improvement in psoriasis control, with significant safety concerns occurring only infrequently [20]

13.4 Contributions of Phase IV Studies to Clinical Cancer Research

When it comes to the improvement of health care, it is critical to evaluate an intervention beyond the controlled research setting. Post-approval phase IV studies play a crucial role in understanding the real benefit of new interventions (drugs or devices) in large-scale populations, translating the efficacy seen in the well-controlled environment of clinical trials into real-world effectiveness [12]. Some particular groups, such as children, pregnant women, elderly patients, patients with comorbidities other than the one being studied, or severely ill patients, are often excluded from clinical trials, and in this regard, phase IV trials can help to establish the generalizability of the findings [12, 13]. Nevertheless, some phase IV studies in oncology utilize the same eligibility criteria as their phase III counterparts, as, for example, the Aflibercept Expanded Access Program in the second-line treatment of metastatic colorectal cancer [21]. In such cases, the external validity of treatment-related safety and efficacy may be compromised.

Moreover, continuous monitoring of interventions is necessary for their lifetime, given the rarity of some adverse events. It is estimated that for an adverse event with a frequency of 1 out of 10,000, it would require 65,000 patients to pick up an excess of three adverse events [11, 12], while phase III trials are only capable of detecting adverse events that occur in up to 1 out of 100 persons [22, 23]. Anecdotal reports of unanticipated cardiovascular events associated with Vioxx® (rofecoxib), an anti-inflammatory medication, and sibutramine, an appetite suppressant, illustrate the importance of post-marketing surveillance in preventing further catastrophic outcomes [24].

Despite the relevance of phase IV trials to clinical practice and despite their loose regulations, there are several barriers to randomized trials of public health interventions with regard to random allocation, control groups, the collection of data, and prospective follow-up [25]. The collection of reliable information and evaluation of data are matters of continuous discussion. There are several limitations in the current systems that identify adverse events, with these systems sometimes being dependent on physicians' suspicions and their willingness to report, a factor that raises concern about the quality of data obtained [12]. Furthermore, the follow-up of participants in a non-interventional study may be less thorough than that in controlled trials [24]. In terms of cost-effectiveness, pharmacoeconomic studies often face the challenge of translating frequently used surrogate measures into long-term outcomes, as well as facing challenges in their adjustment to simulation models of economic performance [12].

Phase IV trials also lack clear and well-defined regulations. In 2001, the FDA made it compulsory to carry out post-marketing studies (or commitment studies) for new drug applications, but a report from 2006 exposed the fragility of this system: of 1231 commitment studies registered, 65% were still pending, 18% were ongoing, and only 14% were completed [12].

Conclusion

In conclusion, phase IV studies are necessary to better understand the effects of new interventions and their interactions in the real-world setting. Despite the importance of these studies for patient safety and perhaps for the establishment of optimized approaches, there is a lack of guidelines and regulatory criteria.

References

1. Gad SC. Clinical trials handbook. 1st ed. New York: Wiley; 2009. ISBN-13: 978-0471213888; ISBN-10: 0471213888
2. Hill TP. Conducting phase IV clinical studies: a moral imperative? Ecancermedicalscience. 2012;6(PMCID: PMC3493021):276.
3. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. JAMA. 2002;287(17):2215–20. PMID: 11980521
4. Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. JAMA. 2004;292(21):2643–6. PMID: 15572722
5. https://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/DataStandardsManualmonographs/default.htm. Accessed 16 Apr 2017.
6. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002877.pdf. Accessed 16 Apr 2017.
7. Ferguson L. External validity, generalizability, and knowledge utilization. J Nurs Scholarsh. 2004;36(1):16–22. PMID: 15098414
8. https://www.fda.gov/NewsEvents/PublicHealthFocus/ExpandedAccessCompassionateUse/. Accessed 26 Apr 2017.
9. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM351261.pdf. Accessed 26 Apr 2017.
10. Iudicello A, Alberghini L, Benini G, Mosconi P. Expanded Access Programme: looking for a common definition. Trials. 2016;17:21.
11. Brower A. Phase 4 research grows despite lack of FDA oversight. Biotechnol Healthc. 2007;4(5):16–22.
12. Glasser SP, Salas M, Delzell E. Importance and challenges of studying marketed drugs: what is a phase IV study? Common clinical research designs, registries, and self-reporting systems. J Clin Pharmacol. 2007;47(9):1074–86.
13. Suvarna V, Phase IV. Of drug development. Perspect Clin Res. 2010;1(2):57–60. PMID: 21829783
14. Moreno C, Montillo M, Panayiotidis P, Dimou M, Bloor A, Dupuis J, Schuh A, Norin S, Geisler C, Hillmen P, Doubek M, Trněný M, Obrtlikova P, Laurenti L, Stilgenbauer S, Smolej L, Ghia P, Cymbalista F, Jaeger U, Stamatopoulos K, Stavroyianni N, Carrington P, Zouabi H, Leblond V, Gomez Garcia JC, Rubio M, Marasca R, Musuraca G, Rigacci L, Farina L, Paolini R, Pospisilova S, Kimby E, Bradley C, Montserrat E. Ofatumumab in poor-prognosis chronic lymphocytic leukemia: a Phase IV, non-interventional, observational study from the European Research Initiative on Chronic Lymphocytic Leukemia. Haematologica. 2015;100(4):511–6.
15. Crino L, Dansin E, Garrido P, et al. Safety and efficacy of first-line bevacizumab-based therapy in advanced non- squamous non-small-cell lung cancer (SAiL, MO19390): a phase 4 study. Lancet Oncol. 2010;11:733–40.
16. Komatsu Y, Ohki E, Ueno N, Yoshida A, Toyoshima Y, Ueda E, Houzawa H, Togo K, Nishida T. Safety, efficacy and prognostic analyses of sunitinib in the post-marketing surveillance study of Japanese patients with gastrointestinal stromal tumor. Jpn J Clin Oncol. 2015;45(11):1016–22.

17. Edwards BJ, Raisch DW, Saraykar SS, Sun M, Hammel JA, Tran HT, Wehr N, Arabyat R, West DP. Hepatotoxicity with Vismodegib: an MD Anderson Cancer Center and Research on Adverse Drug Events and Reports Project. Drugs R D. 2017;17(1):211–8. PMID: 28063021
18. Patel MM, López-Collada VR, Bulhões MM, De Oliveira LH, Bautista Márquez A, Flannery B, Esparza-Aguilar M, Montenegro Renoiner EI, Luna-Cruz ME, Sato HK, Hernández-Hernández Ldel C, Toledo-Cortina G, Cerón-Rodríguez M, Osnaya-Romero N, Martínez-Alcazar M, Aguinaga-Villasenor RG, Plascencia-Hernández A, Fojaco-González F, Hernández-Peredo Rezk G, Gutierrez-Ramírez SF, Dorame-Castillo R, Tinajero-Pizano R, Mercado-Villegas B, Barbosa MR, Maluf EM, Ferreira LB, de Carvalho FM, dos Santos AR, Cesar ED, de Oliveira ME, Silva CL, de Los Angeles Cortes M, Ruiz Matus C, Tate J, Gargiullo P, Parashar UD. Intussusception risk and health benefits of rotavirus vaccination in Mexico and Brazil. N Engl J Med. 2011;364(24):2283–92. https://doi.org/10.1056/NEJMoa1012952. PMID: 21675888.
19. Durán CE, Christiaens T, Acosta Á, Vander Stichele R. Systematic review of cross-national drug utilization studies in Latin America: methods and comparability. Pharmacoepidemiol Drug Saf. 2016;25(1):16–25. PMID: 26486230
20. Driessen RJ, Boezeman JB, van de Kerkhof PC, de Jong EM. Three-year registry data on biological treatment for psoriasis: the influence of patient characteristics on treatment outcome. Br J Dermatol. 2009;160(3):670–5. PMID: 19210502
21. Riechelmann R, Srimuninnimit V, Kavan P, Di Bartolomeo M, Maiello E, Cicin I, Kröning H, Garcia-Alfonso P, Chau I, Fernández-Martos C, Ter-Ovanesov M, Peeters M, Picard P, Bordonaro R. Aflibercept plus FOLFIRI for 2nd line treatment of metastatic colorectal cancer (mCRC): long-term safety observation from the global aflibercept safety and quality-of-life (QoL) program (ASQoP). Ann Oncol. 2016;27(Suppl 6):552P. https://doi.org/10.1093/annonc/mdw370.100.
22. Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. BMJ. 1995;311(7005):619–20. PubMed: 7663258
23. Castle WM, Lewis JA. Postmarketing surveillance of adverse drug reactions. Br Med J (Clin Res Ed). 1984;288:1458–9.
24. James WP, Caterson ID, Coutinho W, et al. Effect of sibutramine on cardiovascular outcomes in overweight and obese subjects. N Engl J Med. 2010;363(10):905–17. PubMed: 20818901
25. Bonell CP, Hargreaves J, Cousens S, et al. Alternatives to randomization in the evaluation of public health interventions: design challenges and solutions. J Epidemiol Community Health. 2011;65(7):582.

Identifying Bias in Clinical Cancer Research

14

Francisco Emilio Vera-Badillo and Rachel P. Riechelmann

14.1 Introduction

Evidence-based clinical medicine relies on the publication of high-quality data to determine the standards of patient care [1]. While phase III trials set the basis for determining the best treatment options for patients, uncontrolled phase II trials, prospective observational cohorts, and retrospective and case-controlled studies are also important as hypothesis-generating studies and for providing ancillary information on the efficacy and toxicity of cancer-directed therapies. Hence, the quality of such studies is important for further phase III development and for helping to make clinical decisions. Bias, a distorted form of gathering, analyzing, and interpreting scientific data, can occur in any of such studies and can potentially invalidate research findings.

Bias can also be present at the time of study reporting. The accurate presentation of the results of a randomized controlled trial (RCT) is the cornerstone of the dissemination of the results and their implementation in clinical practice [2]. Scientific articles are not simply reports of facts, and authors have many opportunities to consciously or subconsciously shape the impression of their results for readers; bias in

F. E. Vera-Badillo, M.D., M.Sc. (✉)
Canadian Cancer Trials Group, Queen's University, Kingston, ON, Canada

Department of Oncology, Queen's University, Kingston, ON, Canada

Centro Universitario Contra el Cancer, Hospital Universitario Dr. Jose E. Gonzalez,
Universidad Autonoma de Nuevo Leon, Monterrey, Mexico

R. P. Riechelmann, M.D., Ph.D.
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_14

use of the language reporting outcomes (i.e., spin) can distort the interpretation of results and mislead readers. The use of these techniques can result from ignorance of the scientific issues, unconscious bias, or willful intent to deceive [3]; favorable results are often highlighted while unfavorable data may be suppressed [4]. Also, while appropriate authorship establishes accountability, responsibility, and credit for the scientific information reported in biomedical publications, misappropriation of authorship undermines the integrity of the authorship system and can be associated with other types of bias (e.g., selection bias) [5].

Given all these issues, great efforts have been made by the scientific community and society in general to avoid the introduction of bias in clinical cancer research. Such efforts include the peer-review system for research publications, registration of trials in public databases to avoid publication bias, publication of research protocols along with the main study results to ensure transparency, and check-list criteria for reporting the results of RCTs. More and more oncology journals require authors to detail the methodology, statistical planning, and analyses of the studies, as well as providing a standardized declaration of financial conflicts of interest of the authors, among other items. In this chapter we will discuss the most common biases encountered in the methodology and reporting of clinical oncology research (Table 14.1).

Table 14.1 The most common types of bias in clinical cancer research and their characteristics

Type of bias		Definition	Study designs prone to bias	Potential solutions
Selection bias	Overall	Partial selection of subjects who are different from the general population in terms of prognostic (and possibly predictive) factors	All types	Randomization; selection of consecutive subjects; multicenter studies
	Immortal time bias	A type of selection bias that, when introduced, incorrectly attributes better survival to the experimental group owing to "immortal time"	Prospective and retrospective cohorts	Adjust analyses for "immortal time"
Ascertainment bias		Occurs when researchers are aware of which treatment patients are receiving. This bias can arise during data collection and/or during analyses	Randomized clinical trials	Double (or triple) blinding; use of matched placebo
Informative bias		Occurs during data collection, when there are systematic errors in the classification of data or inaccuracy in obtaining the data	Any type	Pay attention and pre-determine definitions of variables

Table 14.1 (continued)

Type of bias		Definition	Study designs prone to bias	Potential solutions
	Recall bias	Imprecise memory of patients about their past exposure of interest	Surveys with patients providing the information on exposure	Use accurate sources and instruments with multiple forms to assess the main question to improve precision
	Interviewer bias	Interviewer influences patients' answers, directing them according to a predetermined hypothesis	Surveys, interviews, patient-reported outcome studies	Impartiality and exerting no interference on patients' answers and perceptions
	Regression dilution bias	When investigators do not consider the phenomenon of regression to the mean	Patient-reported outcome studies	Utilize successive measurements of the study variable
	Lead-time bias	The disease is detected earlier than it would otherwise have been detected, owing to the diagnostic test, regardless of its influence on survival	Screening/diagnostic studies	Randomization; be aware of a high risk of contamination from the screening test in the control arm
	Length-time bias	Indolent tumors are more likely to be detected during the screening interval than aggressive tumors, falsely inflating the survival of patients receiving screening	Screening/diagnostic studies	Randomization; be aware of a high risk of contamination from the screening test in the control arm
Measurement bias		Systematic imprecise assessment of a variable	Any type	Cautiously plan the definition of each study variable, preferably using standard definitions; calibrate machines; use validated patient-reported outcome instruments
Bias in reporting	Spin	The use of reporting strategies to highlight that a treatment is beneficial (e.g., emphasizing secondary endpoint findings), despite not presenting statistically significant results	Any type of study where statistical inferences are made; commonly in randomized clinical trials	Pre-define the primary endpoint and parameters to determine a positive result; do not overestimate secondary endpoint results
	Under-reporting of harms	Systematic hiding of information on harms associated with the therapeutic intervention	Any type, but more detrimental in randomized clinical trials	Transparency in reporting data about toxicity/health risks
	Publication bias	Positive trials tend to be published earlier and in high-impact journals	Randomized clinical trials	Register all clinical trials in public databases; publish your research results, even if they are negative

14.2 Definition of Bias and Its Impact on Clinical Cancer Research

Bias is defined as any tendency that prevents the unprejudiced consideration of a question, and it can be intentional or unintentional. In clinical research, bias occurs when systematic errors are introduced into sampling, analyses, interpretation, or reporting by selecting or encouraging one outcome or answer over others [6]. For example, bias can occur when there is inconsistency between the statistical planning and results [7, 8], if endpoints are changed during the course of a clinical trial (usually to allow the reporting of a positive result) [9], or if toxicity is not clearly reported [10]. In study design, outcomes should be defined clearly prior to study implementation; data collection methods should be standardized; and study personnel should be blinded, if possible. During a clinical trial, bias can arise from errors in capturing data, and from the misclassification of exposure or outcome. Unlike random error, which results from sampling variability and which decreases as sample size increases, bias is independent of both sample size and statistical significance. Bias can cause a perceived association, which may be directly opposite of the true association. Importantly, bias is different from confounding, meaning that bias creates a false association, while a confounding factor reports a true, but incorrect, association. A classical example is the incorrect association between coffee drinking and lung cancer and tobacco; while coffee consumption is associated with an increased risk of lung cancer, it is a confounding factor, because people who drink more coffee are more likely to also smoke [11].

After a trial is concluded bias can still occur during data analysis or during the publication process. Citation bias may occur because researchers and trial sponsors may be unwilling to publish unfavorable results, believing that such findings may reflect negatively on the efficacy of their product. Thus, positive results are more likely to be published than negative results [12]. Bias after a trial has been completed is also quite critical because it may distort the correct the interpretation of study results.

Consequently, bias in any of these steps of clinical cancer research directly affects how a drug or intervention is introduced into clinical practice and may ultimately influence treatment decisions and patient care. Therefore the adequate understanding of bias in cancer research in vital for readers of oncology literature.

14.3 Bias in the Design and Analyses of Studies

The production of scientific data is a serious business. Every study, regardless of its type or design, should be carefully planned *before* initiation. Thorough deliberation and brainstorming must take place to determine the study endpoints, its design and population, intervention/s, sample size, and statistical planning for analyses. Meticulous planning of all these topics ensures significant control of bias. However, some biases may still occur and authors have to be aware of them so as to critically appraise their study results. Indeed, authors should be impartial when interpreting their own studies. Likewise, readers have to be attentive to potential bias when interpreting the scientific literature. As well as the careful consideration of the study design and analyses, the use of some methodological and statistical interventions may help in controlling for bias. There are many possibilities of bias in research; the most common biases identified in the design and analyses of clinical oncology studies, as well as their potential solutions, are discussed below.

14.3.1 Selection Bias

Selection bias is probably the most well-known and common type of bias encountered in clinical cancer research. This bias reflects the partial selection of subjects who are different from the general population in terms of prognostic (and eventually predictive) factors, such as, for example, Eastern Cooperative Oncology Group (ECOG) status, cancer stage, and age. The main problem with selection bias is that it directly influences the study results, potentially leading to under- or over-estimated findings, thus compromising external validity. While retrospective cohorts and single-arm clinical trials are particularly prone to selection bias because of the lack of control groups, randomization minimizes imbalances of known, and most importantly, unknown factors. Nevertheless, any study design is threatened by selection bias. Randomized trials in oncology tend to have very strict eligibility criteria (to ensure internal validity) where, for example, only patients with good organ function, ECOG status 0 or 1, and without brain metastases are enrolled. Ethnic minorities and elderly cancer patients have been consistently underrepresented in phase III clinical trials. Finally, patients who are willing to be enrolled in clinical trials may be different from those who decline participation; for instance, with respect to compliance with study procedures and adherence to treatment. All these issues lead to selection bias because they undermine the generalizability of phase III results to cancer patients treated in the community. Phase IV studies and large population database studies are vital tools for evaluating the safety and efficacy of cancer treatments in "real-world" patients.

Selection bias is ubiquitous, as it can be present from study conception (e.g., biased research questions, determination of inclusion and exclusion criteria), in data collection, and in analyses and reporting. To minimize selection bias when defining the eligibility criteria, investigators have to balance internal vs. external validity, enroll consecutive rather than "best candidate" patients, invest in collaborative multicenter studies, provide international standard ancillary medical care to patients, etc. During data gathering for randomized trials, allocation concealment has to be ensured; also, reporting a flowchart of how many patients were eligible and approached, how many accepted/declined participation, and the number of patients analyzed, informs to what extent the analyzed population was narrowed down from the initially eligible subjects, implying selection bias.

Many different forms of selection bias may arise at the time of data analysis. Sufficient follow-up time is important for events (e.g., progression or death) to arise, because short follow-up may underestimate the incidence of events (e.g., cancer progression) or even compromise the study results owing to lack of power. In contrast, participants need to be followed up to the end of the study to avoid losses of follow-up and censoring. Because oncology patients who drop out tend to be sicker than those who remain, a high rate of attrition (attrition bias) may also inflate the study results (e.g., in regard to quality of life); it is also tricky when the attrition rates are different between study groups. Another common example of bias occurs in the comparison of survival outcomes between responders and non-responders. This sort of comparison is biased because patients have to live long enough to be evaluable for response. Landmark analysis is a statistical technique used in this situation, where a fixed time point after treatment initiation is chosen for conducting the analysis of survival according to response. Another common strategy to control for selection bias during analyses is to perform an intention-to-treat analysis (ITT). ITT analyses can be performed in both retrospective and prospective studies (not only in randomized phase III trials). An ITT analysis is an analysis of all patients who were allocated to the intentional intervention, regardless of whether they have crossed over to the other arm, have been lost to follow-up, or were censored. This strategy permits the evaluation, instead of the exclusion, of patients with poor prognosis who may not be fit to have computerized tomography performed, and it permits the assessment of toxic therapies where patients may die from adverse events right after the first doses, before response evaluation is done. Per-protocol analysis, in contrast, selects which patients should be analyzed, e.g., patients who have completed two cycles of chemotherapy or who have undergone surgical resection. ITT is more conservative than per-protocol analysis and thus, is preferable when reporting the results of superiority studies. It is also crucial that researchers describe the number of patients *planned* to be analyzed and the actual number that *was* analyzed for each study endpoint, so that readers can critically interpret the results, in terms of loss of data and potential selection bias.

Stage migration can also be associated with selection bias. Stage migration relates to patients treated in different periods of time who are staged by imaging tests with distinct accuracy that may produce different outcomes. For example, with more advanced imaging techniques, more stage IV cancer patients are being diagnosed with oligometastatic disease, rather than stage III disease (defined by less sensitive imaging methods), which may erroneously suggest that metastatic patients are living longer as a consequence of some new intervention. Stage migration is a very important concept that has to be taken into account when interpreting and contextualizing the results of studies conducted in different decades, trials that used different staging methods at different times, and treatment with modern vs. old radiation techniques, etc.

14.3.1.1 Immortal Time Bias

Immortal time bias is a type of selection bias that is introduced in prospective studies, incorrectly attributing better survival to the experimental group. This may happen because the period of “immortal time”, e.g., the time interval during follow-up where death or other survival endpoints are not considered owing to the exposure definition in the study design, is either incorrectly attributed to the experimental group or is excluded from analysis because the start of follow-up for the experimental group was later than that of the control group [13]. In both scenarios, there is an inflation of the survival time of the experimental therapy. For example, immortal time bias may explain why, in a population cohort study, neoadjuvant chemotherapy improved survival in resectable pancreatic cancer patients compared with survival in patients with upfront surgical resection followed by adjuvant treatment [14]. In this example, patients in the neoadjuvant group must have lived long enough to undergo surgery, which implies an immortal time related to the duration of neoadjuvant chemotherapy; patients who die before surgery are excluded. Simultaneously, patients in the upfront surgical group must live until the end of adjuvant therapy to be included in the analyses; here the immortal time is the duration of adjuvant therapy. In another example, metformin use ceased to impact positively on the survival of diabetic patients with pancreatic cancer, after correcting for the timing of initiation and duration of exposure to metformin [14].

14.3.2 Observation or Ascertainment Bias

Ascertainment bias occurs when the results of an RCT are distorted because researchers are aware of which treatment patients are receiving. This bias, which can arise during data collection and/or analyses, may introduce misleading attributions of drug-related adverse events or objective response evaluation because it prevents impartial judgment. Indeed, studies without proper allocation concealment tend to favor the experimental interventions [15]. For example, if an investigator knows his/her patient is receiving a targeted agent instead of placebo exclusively, he/she may incorrectly classify an adverse event of fatigue as being related to the new drug rather than the event being a consequence of cancer progression. Ideally, ascertainment bias could be greatly minimized if all people involved in a trial are blinded (nurses, pharmacists, investigators, patients, etc.). However, this is not logistically easy and most randomized trials impose double-blinding, meaning the blinding of investigators and patients. The best way to achieve blinding during data capture (e.g., during patient clinical evaluations) is to use a matching (identical to the experimental therapy) placebo. While there is an ethical debate about the use of placebo exclusively in cancer trials, placebos can help in controlling for patients' subjective improvement, called the “placebo effect” [16], and reduce bias on the part of investigators (observer bias).

Observation bias can also originate from the participants, i.e., those *under observation*. The Hawthorne effect [17] is a phenomenon where people who are knowingly observed by others tend to perform better. While this effect has been widely discussed in psychosocial research, it is potentially real in cancer clinical trials or prospective studies of patient-reported outcomes. For example, patients may improve their self-perceived health status, sense of well-being, tolerance to therapy, and even quality of life because they are being watched (or cared for) by investigators and nurses. The Hawthorne effect in RCTs may potentially compromise external validity.

With respect to data analysis, ascertainment bias can occur if statisticians and/or investigators are aware of patients' treatment allocation. This bias can be reduced when analysts deal with treatment groups as codes, so that they are blinded to study groups; such codes should be broken only *after* complete data analyses have been performed.

14.3.3 Informative Bias

This type of bias occurs during data collection and can result from errors in the classification of data (e.g., responders vs. non-responders), different classifications among groups (e.g., time-to-progression being measured differently in two groups because of distinct intervals between imaging tests), or biased methods of obtaining the data. In regard to bias in methods of obtaining the data, the most common types are recall bias, interviewer bias, observer bias (discussed above), and regression dilution bias [18].

Recall bias is associated with patients' inaccurate (or partial) memory of their past exposure of interest. For example, in a case-control study about risk factors for cancer, patients who develop the disease may recall more details about their past medications, dietary habits, and smoking history than the control group of patients without cancer. Interviewer bias may originate in surveys or studies with interviews. Here an interviewer may influence patients' answers, directing them to respond according to the interviewer's preconceived hypothesis; this may be done by "helping" participants to fill out questionnaires, actively asking questions that were supposed to be freely completed by patients, putting more emphasis on certain topics, etc. The regression dilution bias results from not taking into account the natural phenomenon of regression to the mean. This natural law determines that an extreme value measured on its first assessment will likely be less extreme on successive measurements. The regression dilution bias may appear in longitudinal studies that compare baseline with subsequent values of continuous variables. For example, investigators want to conduct a clinical trial to test a new drug for the treatment of diarrhea associated with carcinoid syndrome; they plan to compare the number of bowel movements per day at baseline and after 2 months of treatment. Here each patient will complete a questionnaire about their daily bowel movements. If patients compute the number of bowel movement per day based on the the previous day's information, it may happen that the day before was atypical (extreme value) in terms of carcinoid diarrhea and thus, this information may be inaccurate. To improve preciseness, researchers could collect information about number of bowel movements though a patient diary, kept for, say, a whole week, and determine the mean (or median) number as the baseline value.

14.3.3.1 Lead-Time and Length-Time Biases

Lead-time and length-time biases originate in prospective studies that evaluate screening/diagnostic methods or strategies [19]. In lead-time-bias, survival is magnified as a consequence of detecting the disease *earlier*, irrespective of the intervention's potential to defer death. Hence, a lead time is added to a patient' survival time solely because of an earlier diagnosis—potentially leading to over-diagnosis. The controversy over establishing the prostate-specific antigen (PSA) test as a definitive screening tool for prostate cancer is a typical example of how lead-time bias can falsely overestimate survival. While randomization can control for lead-time bias, it is not easy to randomize patients for cancer screening interventions, given the risk of contamination in the unscreened group.

Length-time bias is similar to lead-time bias, as it also suggests no benefit in screening for diseases. Length-time bias relates to the intervals of screening tests, where more indolent (often asymptomatic) tumors are more likely to be detected through screening, whereas more aggressive cancers will be diagnosed clinically and at a time different from the screening tests. This diagnosis of indolent cancers may incorrectly suggest that the screening intervention improved survival. Screening studies are also subject to over-diagnosis.

14.3.4 Measurement or Instrumental Bias

Measurement bias strikes when there is systematic imprecise assessment of a variable; hence, this bias is linked to the method (or instrument) of measurement used to determine the study endpoints. To achieve a reliable evaluation of study variables researchers need to use standard criteria (e.g., RECIST [Response Evaluation Criteria In Solid Tumors] for objective response evaluation), a validated questionnaire (e.g., the Functional Assessment of Cancer Therapy [FACT] to estimate health-related quality of life), and precise (e.g., calibrated sphygmomanometer to assess blood pressure) measurement instruments or criteria (e.g., definition of cut-offs for determining immunohistochemistry positivity). For example, in retrospective studies, measurement bias may underestimate the real frequency of treatment-induced adverse events, because the toxicity data collected in routine practice may not follow standardized criteria. In clinical trials, the evaluation of response must be done by utilizing the same imaging method for each patient, not allowing patients to be staged with computed tomography (CT) scans at baseline and then have response evaluated by magnetic resonance imaging.

14.4 Bias in Reporting Results

14.4.1 Spin in Reporting Outcomes

Spin, a type of bias, is defined as the use of reporting strategies to highlight that the experimental treatment is beneficial—despite there being a statistically non-significant difference for the primary outcome—or to distract the reader from statistically non-significant results [20]. It is important to recognize the presence of bias and spin in reports of clinical trials, and to evaluate their importance when placing an RCT in context and ascribing a level of credibility to it [21].

Bias in the reporting of outcomes has been explored previously in RCTs with statistically non-significant results for primary outcomes. Boutron [2] et al. performed a study based on the general medicine literature; only studies with a negative primary endpoint were included for consideration, defined as those studies with a primary endpoint that had a *p*-value ≥0.05. Of relevance, the funding source was also recorded at the time of data collection. Evidence of spin was searched for in each section of the article: Abstract Results and Abstract Conclusions; and Results, Discussion, and Conclusions in the main text. Spin was considered to be present when: (1) there was a focus on statistically non-significant results, (2) statistically non-significant results for the primary outcomes were interpreted as showing treatment equivalence or comparable effectiveness; and/or (3) a beneficial effect of the treatment was claimed or emphasized despite the results being statistically non-significant. Seventy-two studies were analyzed. In this study, Boutron et al. reported the presence of spin in the title in 18% of the articles, and in 38% and 58% of the Results and Conclusions sections of the Abstract, respectively. Spin was identified in 29%, 43%, and 50% of the Results, Discussion, and Conclusions sections, respectively, of the main text of the articles. The spin strategies that were used ranged from focusing on within-group comparisons and subgroup analyses in the Results section to a focus, in the Conclusion, of only the beneficial effect of treatment. There were inappropriate claims for equivalence or comparable effectiveness, and claims for efficacy arising from a focus on statistically significant results in non-primary endpoints [2]. Spin was used more commonly to report Conclusions in the Abstract sections than in any other sections of the articles.

More recently, in an analysis of lung cancer clinical trials, Sacher et al. noted that 53% of studies published between 2001 and 2010 were reported as positive, despite not achieving statistical significance in their primary outcome, compared with 24% of studies published between 1991 and 2000 showing such reporting. These inappropriate conclusions were based on improvements seen in secondary trial endpoints, asserting non-inferiority despite the lack of a statistically appropriate non-inferiority design, or recommending further study on the basis of a non-significant trend in the primary outcome [22].

Previously, we reported a decision tree that was used to assess whether the primary endpoint was reported with bias, and whether a secondary endpoint was used to imply benefit of the experimental arm [9]. We considered that all reports using a statistically significant secondary endpoint to highlight the results of a specific trial when the primary endpoint was negative should be regarded as having bias in reporting efficacy. We explored this scenario in two reports: in breast cancer trials, we reported that 59% of 92 trials with a negative primary endpoint used secondary endpoints to suggest benefit of the experimental therapies. A second article on trials reporting the results of phase III RCTs in the field of medical oncology was assessed, showing that, in 107 of 200 RCTs, 50 (47%) used biased reporting in the Abstract to imply benefit of the experimental treatment, although there were no statistically significant primary endpoint results [23].

14.4.2 Underreporting of Toxicity in Oncology Trials

The reporting of harm is as important as the reporting of efficacy in publications of clinical trials. Both are essential for estimating the ratio of benefit to harm of medical interventions. However, harm is frequently insufficiently reported. Ioannidis and Contopoulos-Ioannidis described the reporting of harms as "in general inadequate" [24].

Reviews have shown that a substantial proportion of clinical trials have suboptimal reporting of harm [25]. The use of guidelines such as the Consolidated Standards of Reporting Trials (CONSORT) can improve the quality of reporting of clinical trials [26], by including the mandatory reporting of toxic side effects of the treatment under evaluation.

Several reviews have shown that a substantial proportion of clinical trials have suboptimal reporting of harm [10]. Pitrou et al. reported results based on 133 studies in general medicine. Reporting of any adverse event in the Abstract, the section that physicians mostly read, was deficient, with 47% to 85% of the studies not reporting adverse events in this section. No information on the severity of adverse events was given in 27% of the studies, and 12% reported only generic statements, with only 16% of the studies describing explicitly the grading of severity. Beyond describing the incidence of adverse effects, statistical analysis to obtain objective conclusions was not very frequent, and only 47% of the analyzed studies used at least one statistical test to compare safety data. Reporting of toxicity in Tables was not present in all publications: 32% of the articles did not include a Table or Figure describing toxicity [20].

Of great importance, and under-recognized, is the fact that phase III trials are usually underpowered to detect differences in harms between the study arms, so that the commonly used phrase "no significant differences were found" is misleading [25]. The lack of prominence given to side effects is such that, in a study by Seruga et al. it was reported that 39% of potentially serious adverse drug reactions were not described in phase III cancer trials [27]. RCTs are well known to have insufficient statistical power to assess safety outcomes. Tsang et al. [28] showed that, in a sample of RCTs, the power to detect a statistically significant difference in serious adverse events yielded values ranging from only 0.07 to 0.37. As highlighted by Ioannidis et al.: "We must no longer accept confusing lists of non-comparable percentages of adverse events for clinical or scientific purposes. The lists can needlessly alarm patients and physicians or invite dismissal of real medication hazards" [29].

The above studies highlight the need for improving the reporting of harm-related results in clinical trials. Despite the extension of the CONSORT statement to include harm-related data, efforts are still needed to describe safety results with accuracy in the reporting of RCTs and to standardize practices for reporting. In an attempt to standardize the evaluation of toxicity reporting, we created a hierarchy scale to categorize the quality of reporting [9]. This scale ranges from 1 (excellent) to 7 (very poor) to indicate whether reporting of grade 3 and 4 toxicities occurred in the concluding statement of the Abstract, elsewhere in the Abstract, in the Results section of the article, only in a Table, or not at all, with lower scores if these toxicities were also included in the Discussion section of the paper. We defined the reporting of grade 3 and 4 toxicities as poor if they were not mentioned in the Abstract (scale of 5–7 in our hierarchy), and good (scale 1–2) if they were mentioned in the concluding statement of the Abstract. When there were no statistically significant differences in toxicity, a general statement in the Abstract was deemed to be sufficient; when statistically significant differences were seen, it was expected that they would be reported in the Abstract. Utilizing the aforementioned hierarchy scale, in a cohort of 164 trials reporting phase III RTCs in breast cancer, 110 studies met the definition of poor reporting of toxicity, only 32% of articles indicated the frequency of grade 3 and 4 toxicities in the Abstract, and more importantly, there was a statistically significant association between biased reporting of toxicity and observation of a statistically significant difference in the study arms for the primary endpoint (odds ratio [OR] = 2.00, 95% confidence interval [CI] = 1.02–3.94, $p = 0.044$), and statistically significant studies underreported toxicity [9]. We also analyzed the reporting of harms/adverse events in RCTs in general medical oncology. Of 200 evaluable trials, 37 (18.5%) did not provide a description of toxicity in the Abstract and met the criterion for underreporting, and only 48 (24%) articles summarized toxicity in the concluding statement(s) of the Abstract. Again, there was a statistically significant association between underreporting of toxicity and the observation of a statistically significant difference between the arms for the primary endpoint; studies with a positive result for the primary endpoint were more likely to underreport toxicity (OR = 4.76; 95% CI = 2.15–8.44; $p < 0.001$) compared with the studies with a negative result [23].

14.4.3 Consistency in the Reporting of Primary Endpoints

Selection of endpoints or outcome measures is another concern. Although in 2004 the International Committee of Medical Journal Editors (ICMJE) published guidelines for the mandatory registration of clinical trials [30] and in 2007 the Surgical Journal Editors Group followed these recommendations, consistency between a clinical trial registry and the final article in the reporting of primary and secondary endpoints of surgical RCTs is poor, with one study reporting that 45% of articles had omissions, additions, changes in definition, and downgrading or upgrading of outcomes [6]. Another study showed similar results, with 49% discrepancies in the reporting of primary outcomes in general surgery [31].

Another substantial problem is that primary outcomes are not always clearly identified or defined. Having the primary endpoint objectively defined is crucial for study interpretation and, consequently, for decisions about patient care.

Also important is publication bias, which refers to the selective reporting of trials with apparently beneficial results, and this, together with other strategies, such as changing the primary endpoint from a negative one to a new positive endpoint, affects the credibility of reported studies [32, 33]. These and other factors have stimulated the development of trial registries, which are publically available, as a key tool for reducing bias in reporting [34, 35]. In 2005 the ICMJE initiated a policy requiring investigators to deposit information about trial design in an accepted clinical trials registry before the onset of patient enrollment [36], thereby improving transparency. Registries need to meet minimum criteria and the editors of most high-impact journals have established such registration as a requirement for publication [37]. Information in clinical trial registries should reflect precisely the protocol used in the clinical trial, but there are no reports confirming that data in the registry do accurately reflect the protocol; discrepancies between endpoints reported in the registry and those finally reported in the published article have been described in up to 49% of trials [31]. Clarity in registration is required to determine whether there was a deviation from the protocol [38].

Conclusions

Bias is a real threat to study validity and accuracy and, therefore, it must be avoided as much as possible by investigators. Because bias can occur with any type of trial and at any time of research development, awareness of the most common types of bias is crucial for professionals working in clinical cancer research. In clinical oncology research specifically, the most common types of bias are selection bias, informative bias, ascertainment bias, and measurement bias—with all their nuances and ramifications—and spin bias. We strongly believe that there should be more discussions on study interpretation and bias by the scientific community to boost critical thinking in order to ultimately improve the care of cancer patients.

References

1. Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. PLoS Med. 2008;5(11):e217. discussion e217
2. Boutron I, et al. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303(20):2058–64.
3. Fletcher RH, Black B. "Spin" in scientific writing: scientific mischief and legal jeopardy. Med Law. 2007;26(3):511–25.
4. Chan AW. Bias, spin, and misreporting: time for full access to trial protocols and results. PLoS Med. 2008;5(11):e230.
5. Flanagin A, et al. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. JAMA. 1998;280(3):222–4.
6. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. Plast Reconstr Surg. 2010;126(2):619–25.
7. Marco CA, Larkin GL. Research ethics: ethical issues of data reporting and the quest for authenticity. Acad Emerg Med. 2000;7(6):691–4.
8. Hrobjartsson A, Gotzsche PC. Powerful spin in the conclusion of Wampold et al.'s re-analysis of placebo versus no-treatment trials despite similar results as in original review. J Clin Psychol. 2007;63(4):373–7.
9. Vera-Badillo FE, et al. Bias in reporting of end points of efficacy and toxicity in randomized, clinical trials for women with breast cancer. Ann Oncol. 2013;24(5):1238–44.
10. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 1997;315(7109):640–5.
11. Galarraga V, Boffetta P. Coffee drinking and risk of lung cancer-a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2016;25(6):951–7. Epub 2016 Mar 28. https://doi.org/10.1158/1055-9965.EPI-15-0727.

12. Tam VC, et al. Compendium of unpublished phase III trials in oncology: characteristics and impact on clinical practice. J Clin Oncol. 2011;29(23):3133–9.
13. Zhou Z, et al. Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: a comparison of methods. Am J Epidemiol. 2005;162(10):1016–23.
14. Chaiteerakij R, et al. Metformin use and survival of patients with pancreatic cancer: a cautionary lesson. J Clin Oncol. 2016;34(16):1898–904.
15. Pildal J, et al. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol. 2007;36(4):847–57.
16. Chvetzoff G, Tannock IF. Placebo effects in oncology. J Natl Cancer Inst. 2003;95(1):19–29.
17. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol. 2014;67(3):267–77.
18. Tripepi G, et al. Bias in clinical research. Kidney Int. 2008;73(2):148–53.
19. Croswell JM, Ransohoff DF, Kramer BS. Principles of cancer screening: lessons from history and study design issues. Semin Oncol. 2010;37(3):202–15.
20. Pitrou I, et al. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med. 2009;169(19):1756–61.
21. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol. 2007;60(4):324–9.
22. Sacher AG, Le LW, Leighl NB. Shifting patterns in the interpretation of phase III clinical trial outcomes in advanced non-small-cell lung cancer: the bar is dropping. J Clin Oncol. 2014;32(14):1407–11.
23. Vera-Badillo FE, et al. Bias in reporting of randomised clinical trials in oncology. Eur J Cancer. 2016;61:29–35.
24. Ioannidis JP, Contopoulos-Ioannidis DG. Reporting of safety data from randomised trials. Lancet. 1998;352(9142):1752–3.
25. Ioannidis JP. Adverse events in randomized trials: neglected, restricted, distorted, and silenced. Arch Intern Med. 2009;169(19):1737–9.
26. Ioannidis JP, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med. 2004;141(10):781–8.
27. Seruga B, et al. Reporting of serious adverse drug reactions of targeted anticancer agents in pivotal phase III clinical trials. J Clin Oncol. 2011;29(2):174–85.
28. Tsang R, Colley L, Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. J Clin Epidemiol. 2009;62(6):609–16.
29. Ioannidis JP, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. Ann Intern Med. 2006;144(4):298–300.
30. Reed DA, et al. Association between funding and quality of published medical education research. JAMA. 2007;298(9):1002–9.
31. Hannink G, Gooszen HG, Rovers MM. Comparison of registered and published primary outcomes in randomized clinical trials of surgical interventions. Ann Surg. 2013;257(5):818–23.
32. Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol. 1986;4(10):1529–41.
33. Kirkham JJ, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010;340:c365.
34. International Collaborative Group on Clinical Trial Registries. Position paper and consensus recommendations on clinical trial registries. Ad Hoc Working Party of the International Collaborative Group on Clinical Trials Registries. Clin Trials Metaanal. 1993;28(4–5):255–66.
35. Dickersin K, Rennie D. Registering clinical trials. JAMA. 2003;290(4):516–23.
36. International Committee of Medical Journal Editors. Clinical Trial Registration. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html. Accessed 19 Feb 2017.
37. Laine C, et al. Clinical trial registration—looking back and moving ahead. N Engl J Med. 2007;356(26):2734–6.
38. Zarin DA, Tse T. Trust but verify: trial registration and determining fidelity to the protocol. Ann Intern Med. 2013;159(1):65–7.

15 Ethics in Clinical Cancer Research

Rodrigo Santa C. Guindalini, Rachel P. Riechelmann,
and Roberto Jun Arai

15.1 Introduction

Cancer care is fraught with various ethical issues. We may find frequent dilemmas in providing access to care and treatments, palliative and end-of-life care, and treating vulnerable populations. These issues often concern any physician in their routine setting. The application of evidence-based medicine (EBM) has proven to be fundamental in the twenty-first century. The clinical data derived from rigorous research protocols to support EBM has moved towards a high level of complexity in order to achieve the best level of evidence. The pursuit to retrieve organized data, however, intersects with routine medical care. To accommodate significant advances in the area of precision medicine and to streamline the drug development process, newer and even more complex clinical trial design approaches have emerged, such as adaptive, basket, and umbrella trials [1].

To navigate cancer research ethically, strategies should be constructed carefully, because some new clinical experiments tend to move close to unacceptable ethical boundaries. Cancer research develops in the setting of the risks of a life-threatening disease, and this scenario may lead to misinterpretation in over-emphasizing

R. S. C. Guindalini, M.D., Ph.D. (✉)
CLION, CAM Group, Salvador, BA, Brazil

Department of Radiology and Oncology, State of São Paulo Cancer Institute, Faculty of Medicine, University of São Paulo, São Paulo, SP, Brazil
e-mail: rodrigoscg@gmail.com

R. P. Riechelmann, M.D., Ph.D.
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil
e-mail: rachelri2005@gmail.com; Rachel.riechelmann@accamargo.org.br

R. J. Arai, M.Sc., Ph.D.
Clinical Research Unit, State of São Paulo Cancer Institute, Faculty of Medicine, University of São Paulo, São Paulo, SP, Brazil
e-mail: roberto.arai@hc.fm.usp.br

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_15

benefits over risks. Historically, the famous 1747 scurvy trial conducted by James Lind contained most elements of a controlled trial and provided meaningful results, but the intervention lacked any potential for toxicity [2]. Over the years, scandals, such as the elixir of sulfanilamide that killed 107 people in the United States in 1937, have raised concerns regarding what can be offered to the population. The pharmaceutical company involved in that incident was not undertaking any illegal activity in marketing the product [3]. In response to deliberate abuses, of which the most infamous were the torture of and experimentation on Jewish people during World War II and the experiments conducted on black people in Tuskegee, Alabama [4], there have been ethical advances in human protection, such as the Declaration of Helsinki, which has been discussed and updated constantly [5]. Other ethics codes have become the basis of clinical research regulations: the Nuremberg Code (1947); the Belmont Report (1979); the International Conference of Harmonization—Good Clinical Practice (1996), and the Council for International Organization of Medical Sciences (2002). Ethics committees have been created, based on ethical and regulatory guidelines, to critically review projects and to obtain consensus on research validity.

Participation in a clinical trial entails an essentially unknown (and often unpredictable) level of risk. Indeed, the equipoise principle determines that a participant should be enrolled in a clinical trial only when there is uncertainty. This principle aims to avoid the conduct of biased clinical trials, as for example, when a known inferior control intervention is chosen to increase the chances of achieving a positive result, with false claims that the experimental treatment is better [6].

Information that makes a participant's consent valid is generally thought to include an understanding of the risks and benefits of the intervention, understanding and acceptance of the procedures that the participant may undergo, comprehension that participation in the research is voluntary, and an understanding of the research goals. Checking for a participant's comprehension of the consent process is of great importance, because each participant has a particular history, cultural background, and beliefs. Thus, the desire to participate in a clinical trial and the willingness to complete the study are multifaceted. It is critical to address, a priori, individual education and values from the patient's perspective and expectations from the physicians' perspective. Shared decision-making may be defined as "an approach where clinicians and patients share the best available evidence when faced with the task of making decisions, and where patients are supported to consider options, to achieve informed preferences" [7]. This approach should be applied in any decision taken in the clinic, including participation in clinical trials. The informed consent process, however, is not sufficient to make any research ethical. The scientific question should be meaningful and valid; the risks should be minimized and realistically favorable; and patient accrual should be fair [8]. Medical innovation not only creates new ethical concerns, but also adds new considerations and paradigms to long-standing ethics discussions. In this chapter, we will explore some of the major ethical concerns that arise in the course of modern clinical cancer research (i.e., clinical trials), and we will propose recommendations to protect the rights, safety, and welfare of study subjects.

15.2 Early-Phase (Phase 0 and Phase 1) Studies and Therapeutic Expectations

Early-phase oncology clinical trials are an important step in translating basic research into clinical practice. Very different in structure from phase 1 studies, phase 0 trials were introduced by the United States Food and Drug Administration and the Pharmaceutical Research and Manufacturers of America in 2006 [9]. Phase 0 trials can be designed to determine whether a mechanism of action defined in nonclinical models can be achieved in humans; to refine biomarker assays using human tumor and/or surrogate tissue; to provide pharmacokinetic/pharmacodynamic relationship data for an agent prior to phase 1 trials and to select the most promising candidates; and to determine the dose ranges of an experimental drug. This type of study typically involves the administration of a single dose or a short course of micro-doses of a new pharmacological compound [10]. Phase 0 trials are useful because they can compress timelines for the overall development of an anticancer drug or even prematurely halt a drug development program. This would prevent subsequent studies and the unnecessary exposure of volunteers to undesirable drug effects and an unacceptably lower probability of success. Nevertheless, the lack of any potentially direct medical benefit in parallel with the exposure of study subjects to small risks is problematic. Moreover, in routine settings, biodistribution assessments require multiple blood draws and biopsies of relevant organs [11]. This scenario is of great challenge to applied science and ethical responsibility [12]. While studies have reported that cancer trial participants commonly report that altruism contributed to their decision to enroll, it is rare for this to be the primary motivation for study participation. Indeed, the decision to participate frequently stems (at least partially) from the possibility of direct benefit, which is understandable given the lethality of cancer. In early-phase trials and among patients with poor prognoses, altruism is least often the motivator [13, 14]. The common understanding that participants in phase 0 trials should have chances (even minimal) of direct benefits from trial participation or access to the best therapeutic method identified as beneficial during drug development is unsustainable. Because cancer patients who are eligible to enroll in phase 0 trials are end-staged, their chances of receiving direct clinical benefits from proven therapy during the course of drug development are highly limited [15]. Reconceptualization of the investigator-subject relationship and a deep evaluation of the subjects' understanding of phase 0 trials are essential for achieving subject enrollment. Paying modest quantities of money to encourage enrollment in phase 0 trials has been considered by some authors [12]. This idea should have a well-thought-out rationale, especially in developing countries where most participants are in a vulnerable economic condition owing to their low per-capita incomes. Undoubtedly phase 0 trials comprise an important step of the drug development process in clinical cancer research. However, the inclusion of indirect benefits and the potential lack of direct benefits, in terms of tumor control, must be explicit in the consent forms of phase 0 studies.

Phase 1 studies are designed to escalate the dose until toxicity is observed, to determine the recommended safe dose and schedule of an investigational agent in order to move forward to phase 2 trials; phase 1 trials often assess the pharmacokinetics/pharmacodynamics of an experimental therapy and explore the development of relevant biomarkers. The American Society of Clinical Oncology Policy Statement claims that phase 1 trials have the potential to provide clinical benefit, including improved quality of life, positive psychological effects, and the potential for tumor response [16]. Meta-analyses of phase I trials have reported growing rates of objective responses, of 5–11%, a toxicity-related death rate of only 0.5%, and an episode of a grade 4 toxic event in approximately 14% of the participants [17–19]. Even though the efficacy and safety of investigational treatments in early-phase trials are still under evaluation, empirical studies have found that phase 1 oncology trial subjects often report high and unrealistic expectations of personal therapeutic benefit [12, 20]. Unrealistic optimism may cause distortions in risk/benefit assessment, with patients overestimating their prospects of benefit and/or underestimating their susceptibility to the risks. In addition, media 'hype' about laboratory discoveries, as well as cancer centers promoting their medical services by advertising the number of clinical trials they conduct—highlighting that their patients will have access to the newest investigational treatments—can foster therapeutic misconceptions [21, 22]. Such advertising promotes patients' unrealistic hopes of clinical benefit, as well as interfering with the ability of study participants to distinguish between research and standard of care [23]. Therefore, while research participants can, and should, be optimistic about their chances of response to investigational treatments, they must still understand that the treatments are experimental, and that they have potential toxicities and low chances of tumor response [20]. It is challenging to determine whether patients' unrealistic expectations are the result of the patients' optimism, or a result of their lack of understanding during the informed consent process, or both [20, 24]. Although the vast majority of study participants have reported that they understood most of the trial information, fewer than 50% were able to correctly describe the purpose of the phase 1 trial as a dose-determination and safety study [25–29]. These findings raise concern about subjects' misunderstanding of specific information regarding early-phase clinical cancer trials, and make their voluntary decision about whether or not to enroll in the research burdensome.

Contributing to ethical challenges are the complications of conducting research in patients with advanced cancer who have not responded to other types of therapy and have few, if any, remaining treatment choices. Terminally ill cancer patients may seek to enroll or may be actively recruited to participate in early-phase oncology clinical trials while desperately hoping to find something to reverse or delay the course of the disease [30]. Vulnerability, which is closely linked to the disease severity, may affect the informed consent process [25, 31–33]. The debate on the enrollment of such patients focuses on the misinterpretation of risk-benefit ratios, inadequate information disclosure, and subject decision-making capacity [23]. In this context, one can argue that, in high-risk clinical research, the study subjects could be more vulnerable to exploitation and less capable of protecting their own interests, thus requiring special safeguards in this type of research [34, 35]. Nevertheless, after an extensive literature review of almost 10,000 participants in phase 1 oncology trials, Seidenfeld and colleagues concluded that "the demographic and health status (…) are not those of a conventional vulnerable population and suggest little reason to assume that, as a group, they have a compromised ability to understand information or to make informed and voluntary decisions" [36].

15.2.1 Recommendations

15.2.1.1 Improve Subjects' Understanding of Goals of Early-Phase Trials

- During the informed consent process, researchers should clearly discuss with the subjects the objectives of the trial and its potential benefits and risks; in particular the low chance that the subjects will experience clinical improvement.
- Continuing medical education should address the structured training of trialists in communication skills to reduce the frequency of participants' poor understanding of the key concepts of a clinical trial and its objectives [16]. Investigators have to be cautious in properly informing subjects of the low chances of benefit in early-phase trials in cancer, without hampering their hopes.
- To assess and enhance participants' understanding, researchers should use open-ended questions, such as: "Can you tell me in your own words the purpose of phase 1 research?" and "What might be the benefit from this study?" [24].
- If unrealistic optimism becomes apparent during the informed consent process, investigators should attempt to clarify misunderstandings. If there is failure to appreciate relevant information, unrealistic optimism may impair informed consent [20, 24].
- When considering participation in clinical research, especially in phase 1 studies, a shared decision-making approach may be useful when decisions are uncertain. It is important to address, a priori, the patient's individual education and values and the physicians' expectations.

15.2.1.2 Improve the Risk-Benefit Ratio for Patients in Early-Phase Trials

- Researchers should use strategies to facilitate the inclusion of those patients, based on genetic or molecular biomarkers, who are most likely to respond to a specific targeted therapy. Based on a strong biological rationale, enriching the subsets of patients selected according to germline or molecular tumor profiling for matched therapies can improve the efficacy, and potentially the safety, of new cancer-directed experimental drugs [16, 37].
- Researchers and sponsors should put efforts into moving phase 1 trial designs to dose-escalation approaches (e.g., accelerated titration designs and adaptive Bayesian designs) that allow more subjects to receive higher doses of investigational agents that are more likely to result in a therapeutic effect [16, 37].

15.3 Key Ethical Issues in Developing Precision Medicine in Oncology Clinical Trials

Precision medicine is an emerging approach that proposes the customization of disease diagnosis, treatment, and prevention tailored by individual variability in genomic, environment, and lifestyle factors. This concept is rapidly progressing with the recent advances in next-generation sequencing (NGS) technologies (using DNA, RNA, or methylation sequencing), genetics computational solutions for omics data, and large-scale biological databases (such as the gnomAD [38], The Cancer Genome Atlas [39], ClinVar [40], and the Catalogue of Somatic Mutations in Cancer [41]). Each cancer has its own genomic signature and the understanding of the key genomic changes in many types and subtypes of cancer has begun to influence risk assessment, diagnostic categories, and therapeutic strategies in oncology. On the treatment front, the use of predictive biomarkers to select patients for specific molecularly targeted anticancer drugs has established new, more effective and less toxic treatment options. Some examples of these successful approaches are imatinib for chronic myeloid leukemia and trastuzumab for HER2 breast cancer [42].

Precision medicine is powered by patient data. It relies on the integration of clinical, genomic, pathologic, and outcome data, as well as the availability of patient samples. Though these are clear indications of optimism for precision medicine, ethical challenges need to be acknowledged and addressed, particularly with regard to (1) informed consent, (2) privacy/discrimination concerns for patients and their families, and (3) the return of clinically relevant results.

First, for many decades, human biological samples were collected during the course of treatment without gathering any informed consent, or alternatively, the informed consent is now obsolete and inappropriate for use in unforeseen secondary research aims, such as genomic profiling for research purposes [43]. Currently, the lack of individual informed consent imposes an enormous obstacle to the reuse of biological samples [44]. Thus, there is an urgent need to update and standardize patient information and informed consent forms in order to integrate precision medicine into oncology research. The development of a fair informed consent document and the process required for its acquisition, without compromising broad ethical and legal principles, is essential [45]. Many patients have difficulties in understanding the complexity of the information that needs to be covered, such as the tumor's heterogeneity, its molecular evolution, and its relationship to drug response. Not only the content, but also the duration of obtaining informed consent can result in an excessive and undue burden on participants; the extrapolation of the traditional single-gene approach to detailed discussion about each gene being tested may be tremendously time-consuming, leading to information overload; as a consequence, this may interfere with the participant's decision-making capacity [46]. Adding more complexity to the consent process is the duty to warn patients about the potential identification of a genetic variant of unknown significance and incidental findings in medically actionable genes, as well as the potential psychosocial implications of germline and somatic genetic testing [47, 48]. These factors have become more and more common with the widespread use of NGS in routine and clinical research settings.

Second, the maintenance of the privacy and confidentiality of genetic information has been raising ethical concerns in research and ethics communities. On one hand, privacy considerations may restrict researchers from gathering additional information that might give them more insight into their research questions; on the other hand, these considerations are safeguards of anonymity and may prevent unintended consequences and potential risks to participants. A dominant issue of public concern is the potential risk of genetic information being used in ways that could harm people, such as for genetic discrimination by health insurance companies and employers. In the United States, the Genetic Information Nondiscrimination Act of 2008 is supposed to prohibit such use, but there are no similar or specific laws in most countries, particularly in developing countries. Despite researchers making every effort to maintain privacy and confidentiality, according to Neil Savage, "it may not be possible to protect the identity of genomic data" [49]. Privacy should be ensured; however, anonymous data, particularly those shared in public databases, are vulnerable to re-identification [50, 51]. As genome databases are growing and algorithms for comparing data are improving, it is getting easier to link medical histories and other personal information (such as name and ZIP code) to DNA donors.

Finally, as genome and exome sequencing has moved into clinical practice, concerns over unintended/incidental findings and the return of results have emerged. There are several challenges, including the large number of results available, the need to interpret novel mutations that could be functionally deleterious, the reclassification of the variant pathogenicity over time, and the increased number of findings of potential clinical utility. The American College of Medical Genetics and Genomics published, in 2013 [52], and updated in 2016 [47], a report suggesting that, in clinical genomic sequencing, known pathogenic and expected pathogenic genetic variants in 59 medically actionable genes should be returned to the subjects, including when germline testing is done as part of a matched tumor-normal sample pair. Nevertheless, institutional review boards and investigators are raising the question of whether these or comparable clinical recommendations should be extended to research settings. It is important to acknowledge that standards for returning genomic results in a research setting, where investigators are seeking scientific discovery, might differ from the standards of clinical practice, where the clinician's primary duty is to improve the health and wellbeing of the patient [53, 54].

Accumulating evidence reveals that the majority of research participants wish to receive clinically significant individual study results, despite there being a potentially negative emotional impact [55]. Study participants appreciate that the disclosure of study results has the potential for removing uncertainties, promoting discussions within families about risk management and increasing the understanding of a disease and its treatment, as well as reassuring the study participants about their right to know [53, 55]. However, respect for participants' wishes requires taking their preferences seriously, including their right to refuse the return of genetic findings during the informed consent process [56, 57].

15.3.1 Recommendations

15.3.1.1 Improve the Process of Informed Consent

- The time has come to rethink and consider an informed consent model that respects the privacy concerns of participants, but that releases constraints on the utilization of data, making the participants' contribution to research more durable, broader, and efficient.
- Researchers need to develop new variations of informed consent documents, such as tiered or dynamic consent (establishment of ongoing communication between investigators and participants regarding data access) [58], broad consent [59, 60], and open consent (volunteers consent to unrestricted re-disclosure of data, knowing that there is a certain risk of harm to themselves and their relatives and no guarantee of anonymity/privacy/confidentiality) [61].
- The informed consent process should include the option of re-contacting the participant and obtaining re-consent when new medically relevant information becomes available or if further research is being considered.
- Partnering with patients, as for example, in patient advocacy groups, is critical for understanding how to maximize the balance between the ethical/legal regulatory framework, researcher needs, and patient expectations [45].

15.3.1.2 Awareness of Limitations to the Safeguarding of Genetic Privacy and Confidentiality

- Participants need to be fully aware that the linking of distinct databases and data sharing among researchers is intended.
- Investigators need to clearly explain that the full purpose and the extent of further usage of their data cannot be completely foreseen.
- Although the risk of re-identification is small, absolute privacy and confidentiality cannot be guaranteed; thus a certain risk of harm to participants and their family members may potentially exist.
- Continuous efforts have to be made to ensure data safety. For example, remove obvious identifiers from the data sets before sharing the information in public databases and maximize privacy-preserving approaches using new data anonymization methodologies, such as differential privacy and k-anonymity, and modern cryptographic solutions [62].

15.3.1.3 Improve the Communication of Results

- Participants need to be informed about the possibility of incidental findings during the consent process.
- Researchers are not obligated to conduct a deliberate search of a predetermined list of genes not identified in the course of their research or related to their research purpose.
- There is a duty to warn participants, but there is no duty to search for actionable incidental genetic findings. However, there is a duty to return lifesaving genetic information discovered in the course of the research process.

- To respect participants' autonomy—one of the ethical foundations of medicine—participants should have the right to refuse the return of their results. This right should be adequately explained to the participant at the time of consent.
- At the time of consent, if the purpose of the study is dependent on the return of results, the participant must have the opportunity to decline participation.

15.4 Ethical Considerations in Placebo-Controlled Cancer Trials

The most well recognized method of evaluating the efficacy of a new treatment is the double-blind, randomized controlled trial with placebo (placebo-controlled trial—PCT). Beecher was the first to report the placebo effect, noting that, in about 35% of patients with various distinct medical conditions, the conditions could be improved or cured by placebos [63]. From the time of that study, the concept of an intervention activity changed and took into account combined variables: the course of the disease; the specific effect of the intervention; and the nonspecific effects of placebos [64, 65]. Placebo effects have been well documented for the relief of pain [64] and for psychiatric disorders such as anxiety and depression [66]. Responses to treatment in patients receiving placebo are more frequent when the effect is a change in a subjective sensation [67]. In cancer clinical trials with objective responses as the primary endpoint, the use of placebo alone may result in an objective response rate of 2–7%[65]. Although such percentages are considered low, these rates might be overestimated in terms of spontaneous regression or cytokine-mediated regression, or may even reflect measurement errors by radiologists [68]. Analyses of objective responses in patients receiving placebo treatment are still controversial. Also, deleterious effects attributed to placebo may be found in patients who anticipate the potential side effects of the active drug.

In clinical cancer research, placebos are often utilized in randomized registration phase 3 trials. Here a placebo can be used exclusively; in combination with best supportive care, often when there is no active comparator (as for instance, in refractory metastatic solid tumors); or combined with cancer-directed therapies [69, 70]. The advantage of having a placebo arm is that it controls for observation biases in randomized trials (see Chap. 14).

The use of exclusive placebo in randomized trials is permitted in the Declaration of Helsinki, but "extreme care" should be taken. Exclusive placebo can be used when scientifically indicated and when there is no proven effective treatment for the condition under study, or when interrupting treatment poses acceptable risks. However, this idea of "extreme care" has not always been respected. In 1998, trials with azidothymidine (AZT) were being conducted to determine the minimum dose of AZT needed to prevent the vertical transmission of HIV from infected mothers to their unborn children. Volunteers were randomized to various dosage arms and to a placebo arm. The trial format was considered unethical because AZT already had proven efficacy in blocking two-thirds of transmissions of HIV to the fetus [71].

As stated by Emanuel, research must be conducted in a manner that will produce reliable and valid data, and for that new interventions should be tested against the best current proven intervention. Sometimes it will be appropriate to test new interventions against placebo alone, or no treatment, when there is no current proven intervention or, where for compelling and scientifically sound methodological reasons, the use of placebo is necessary to determine the efficacy and/or safety of a new treatment and the patients who receive placebo exclusively, or no treatment, will not be subject to excessive risk or serious irreversible harm [72].

Some authors have justified placebo use in diseases in which worsening is likely to be reversible [73]. In cancer research, however, this idea is unlikely to be acceptable. Indeed, one cancer patient-centered concern commonly includes the understanding of assignment to placebo or no treatment in PCTs [74]. In this regard, accessible information and individualization are some important aspects that influence participants' decisions. If these aspects are not addressed, participation would be compromised and this may result in poor protocol compliance and a higher probability of dropouts [14].

Another important aspect of PCTs is the willingness of patients and physicians to participate in such trials. In fact the use of placebos represents a known barrier to trial enrollment. For example, a survey of nearly 6000 American cancer patients demonstrated that among those who refused trial participation, one-third did so for fear of being administered placebo exclusively [75]. In Brazil, we performed a cross-sectional study of 104 cancer patients and 25 oncologists who were the principal investigators in clinical trials; we asked about their perceptions of a PCT and we found that 41% of patients were not willing to participate in trials with placebos and half of the investigators surveyed objected to recommending a PCT to patients because they "felt uncomfortable to offer no treatment to their patients" [76].

15.4.1 Recommendations

- Specific explanations about the risks and benefits related to the use of placebo should be emphasized during the consent process, including the information that subjects' health status may worsen while on placebo. Clinical equipoise should exist in PCTs and this should be explained to patients in lay terms. We conducted a survey of PCTs in cancer published over a decade and showed that the results of half of the trials were negative, i.e., placebos were not worse, and were certainly less toxic, than the experimental agents under investigation [77].
- To minimize participants' time on placebo, trial withdrawal, early escape, and designs that permit cross-over can be used [78]. However, justifying the use of placebo in cross-over designs should be carefully considered [79]. This is because allowing patients from the placebo group to receive the experimental therapy upon disease progression likely compromises the analysis of overall survival, which may, in turn, preclude the approval of new drugs in countries where gains in survival are a regulatory requirement. Possibly a flexible solution would be to perform imaging tests shortly after treatment initiation, e.g., after 4–6 weeks,

to minimize the use of placebo; this short period might not contaminate the survival analysis when there is significant cross-over.

- To minimize risks associated with the use of placebo in clinical trials, investigators should increase monitoring for deterioration in the subjects' condition and include state-of-the-art palliative care [78, 80].
- The use of unbalanced randomization (2:1 or 3:1 allocation ratio) should be encouraged to keep the population placed on placebo smaller than the number in the active treatment arms [80]; this certainly encourages patients and physicians to accept PCTs.
- A data and safety monitoring board, with interim analyses of study results, should always be considered, with the possibility of early stopping or modifying of the study based on the findings [80]. But again, this possibility has to be carefully considered, because early stopping may lead to trials becoming underpowered to detect differences in overall survival.

15.5 Understanding What Clinical Trial Participation Means

Importantly, a pivotal aspect of clinical research in any area of medicine is to guarantee that patients accept participation voluntarily. The issue is that, to make a voluntary decision, the person has to properly understand what he/she is getting involved in, which implies that the information provided for such a decision is clear, objective, and presented clearly in lay terms. Our perception is that this aspect is far from ideal. The consent forms for clinical trials often consist of more than ten pages, and they contain too much detailed and sometimes useless information for patients, making the whole process of reading tiring and potentially unfruitful. In an attempt to provide detailed information, the consent forms have become burdensome, time-consuming for readers, and confusing. For example, while it is mandatory that consent forms include information about risks, it seems unnecessary to go over the risks of performing basic blood and imaging tests, since these are already part of a cancer patient's life. Mention of such risks may suggest that all interventions included in the trial are experimental, and may wrongly influence patients' decisions to enroll. The other feature of modern consent forms is the great number of technical terms used, which makes the forms hard to comprehend. The list of potential adverse events is often long and described in medical terms, such as neutropenia, hand-foot skin reaction, and increase in QT intervals. This is of particular concern in low-socioeconomic settings. Nowadays most phase 3 trials are global trials that accrue patients from all over the world, including developing countries. In certain countries, like Brazil, the number of illiterate patients and those who have attended only a few years of school is not trivial. These patients are usually treated for their cancers in academic public institutions, where most clinical trials are conducted.

To evaluate the readability and complexity of informed consents for phase 3 trials and the level of education of cancer patients (n=137) who had been enrolled in clinical trials at an academic center in Sao Paulo, Brazil, we performed a transversal study, using widely available software (Flesch Index and Flesch Kincaid Index of readability). We found that understanding the complexity of the consent forms required at least 18 years of education, while half of the patients had attended school for less than 8 years [81]. Making sure patients understand what is at the stake in terms of clinical trial participation is crucial for a transparent and ethical informed consent process.

15.5.1 Recommendations

- There is a need to significantly reformulate the contents of informed consent forms, with less—but more direct and objective—information for patients. Lay terms have to be substituted for technical terms and conventional tests and procedures already performed in the routine practice of oncology should be mentioned, but only as ancillary measures/intervention, without listing all potential risks.
- The language of the consent forms should, if possible, be adapted from—not only translated into—local languages, meaning that the use of certain linguistic terms may make it easier for participants to comprehend scientific terms.
- Enough time must be given for subjects to read and discuss the consent forms with their families or caregivers so that they can reach a voluntary and well-informed decision.

Conclusion

As scientific advances continue, ethical and regulatory challenges requiring wiser updated adaptations will be the tenets of clinical research activities. Recently the International Conference of Harmonization—Good Clinical Practice guidelines were updated, focusing on more complex and globalized studies, with the use of technology applications, including mobile data collection and real-time monitoring of clinical data. The guidelines also recommended the application of a risk-based approach. The conduct of modern research will require critical responsibility on the part of investigators, research promoters, and regulators to guarantee that all terms of the research are within an ethical scope.

In particular, paragraph 8 of the Declaration of Helsinki (2013) emphasizes that while the primary purpose of medical research is to generate new knowledge, this goal can never take precedence over the rights and interests of individual research subjects. In the consent process, critical points still remain in regard to the vulnerability of volunteers; also local factors might be ignored. All efforts should be made to guarantee a valid consent. However, the informed consent process is insufficient to guarantee the ethical conduct of a research program. To guarantee such conduct will require broader monitoring mechanisms, involving regulatory bodies and ethics committees in the validation of the scientific values of research questions, in accrual activities, and in safety concerns. Finally, ethical requirements should not be viewed as clashing with scientific advances. Instead, current understanding of the requirements that make research ethical may be reinterpreted in the light of the modern clinical cancer research scenario and be adapted to boost specific studies and research strategies in the contemporary era.

References

1. Menis J, Hasan B, Besse B. New clinical research strategies in thoracic oncology: clinical trial design, adaptive, basket and umbrella trials, new end-points and new evaluations of response. Eur Respir Rev. 2014;23(133):367–78. PMID 25176973. https://doi.org/10.1183/09059180.00004214.
2. Tröhler U. Lind and scurvy: 1747 to 1795. J R Soc Med. 2005;98(11):519–22. PMID 16260808
3. Wax PM. Elixirs, diluents, and the passage of the 1938 federal food, drug and cosmetic act. Ann Intern Med. 1995;122(6):456–61.
4. Jones DS, Grady C, Lederer SE. "Ethics and clinical research"–The 50th anniversary of beecher's bombshell. N Engl J Med. 2016;374(24):2393–8. PMID 27305197. https://doi.org/10.1056/NEJMms1603756.
5. Fischer BA 4th. A summary of important documents in the field of research ethics. Schizophr Bull. 2006;32(1):69–80. Epub 2005 Sep 28. PMID 16192409
6. Djulbegovic B. The paradox of equipoise: the principle that drives and limits therapeutic discoveries in clinical research. Cancer Control. 2009;16:342–7.
7. Elwyn G, Laitner S, Coulter A, et al. Implementing shared decision making in the NHS. BMJ. 2010;341:c5146.
8. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283:2701–11.
9. Fromer MJ. FDA introduces new phase 0 for clinical trials: some enthusiastic, some skeptical. Oncology Times. 2006;28:18–9.
10. Murgo AJ, Kummar S, Rubinstein L, et al. Designing phase 0 cancer clinical trials. Clin Cancer Res. 2008;14:3675–82.
11. Kummar S, Kinders R, Rubinstein L, et al. Compressing drug development timelines in oncology using phase '0' trials. Nat Rev Cancer. 2007;7:131–9.
12. Abdoler E, Taylor H, Wendler D. The ethics of phase 0 oncology trials. Clin Cancer Res. 2008;14:3692–7.
13. Truong TH, Weeks JC, Cook EF, Joffe S. Altruism among participants in cancer clinical trials. Clin Trials. 2011;8:616–23.
14. Arai RJ, Longo ES, Sponton MH, Del Pilar Estevez Diz M. Bringing a humanistic approach to cancer clinical trials. Ecancermedicalscience. 2017;11:738.
15. Arai RJ, Hoff PM, de Castro G Jr, Stern A. Ethical responsibility of phase 0 trials. Clin Cancer Res. 2009;15:1121. author reply 1121–1122
16. Weber JS, Levit LA, Adamson PC, et al. American Society of Clinical Oncology policy statement update: the critical role of phase I trials in cancer research and treatment. J Clin Oncol. 2015;33:278–84.
17. Horstmann E, McCabe MS, Grochow L, et al. Risks and benefits of phase 1 oncology trials, 1991 through 2002. N Engl J Med. 2005;352:895–904.
18. Roberts TG Jr, Goulart BH, Squitieri L, et al. Trends in the risks and benefits to patients with cancer participating in phase 1 clinical trials. JAMA. 2004;292:2130–40.
19. Decoster G, Stein G, Holdener EE. Responses and toxic deaths in phase I clinical trials. Ann Oncol. 1990;1:175–81.
20. Jansen LA, Mahadevan D, Appelbaum PS, et al. Dispositional optimism and therapeutic expectations in early-phase oncology trials. Cancer. 2016;122:1238–46.

21. Rinaldi A. To hype, or not to(o) hype. Communication of science is often tarnished by sensationalization, for which both scientists and the media are responsible. EMBO Rep. 2012;13:303–7.
22. Vater LB, Donohue JM, Arnold R, et al. What are cancer centers advertising to the public?: a content analysis. Ann Intern Med. 2014;160:813–20.
23. Dresser R. First-in-human trial participants: not a vulnerable population, but vulnerable nonetheless. J Law Med Ethics. 2009;37:38–50.
24. Crites J, Kodish E. Unrealistic optimism and the ethics of phase I cancer research. J Med Ethics. 2013;39:403–6.
25. Daugherty CK, Ratain MJ, Minami H, et al. Study of cohort-specific consent and patient control in phase I cancer trials. J Clin Oncol. 1998;16:2305–12.
26. Itoh K, Sasaki Y, Fujii H, et al. Patients in phase I trials of anti-cancer agents in Japan: motivation, comprehension and expectations. Br J Cancer. 1997;76:107–13.
27. Daugherty C, Ratain MJ, Grochowski E, et al. Perceptions of cancer patients and their physicians involved in phase I trials. J Clin Oncol. 1995;13:1062–72.
28. Schutta KM, Burnett CB. Factors that influence a patient's decision to participate in a phase I cancer clinical trial. Oncol Nurs Forum. 2000;27:1435–8.
29. Hutchison C. Phase I trials in cancer patients: participants' perceptions. Eur J Cancer Care (Engl). 1998;7:15–22.
30. Daugherty CK. Ethical issues in the development of new agents. Investig New Drugs. 1999;17:145–53.
31. Schaeffer MH, Krantz DS, Wichman A, et al. The impact of disease severity on the informed consent process in clinical research. Am J Med. 1996;100:261–8.
32. Daugherty CK. Impact of therapeutic research on informed consent and the ethics of clinical trials: a medical oncology perspective. J Clin Oncol. 1999;17:1601–17.
33. Meropol NJ, Weinfurt KP, Burnett CB, et al. Perceptions of patients and physicians regarding phase I cancer clinical trials: implications for physician-patient communication. J Clin Oncol. 2003;21:2589–96.
34. Jonas H. Philosophical reflections on experimenting with human subjects. Daedalus. 1969;98:219–47.
35. Lipsett MB. On the nature and ethics of phase I clinical trials of cancer chemotherapies. JAMA. 1982;248:941–2.
36. Seidenfeld J, Horstmann E, Emanuel EJ, Grady C. Participants in phase 1 oncology research trials: are they vulnerable? Arch Intern Med. 2008;168:16–20.
37. Wong KM, Capasso A, Eckhardt SG. The changing landscape of phase I trials in oncology. Nat Rev Clin Oncol. 2016;13:106–17.
38. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
39. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat Genet. 2013;45:1113–20.
40. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–5.
41. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.
42. de Bono JS, Ashworth A. Translating cancer research into targeted therapeutics. Nature. 2010;467:543–9.
43. Siu LL, Lawler M, Haussler D, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med. 2016;22:464–71.
44. Rance B, Canuel V, Countouris H, et al. Integrating heterogeneous biomedical data for cancer research: the CARPEM infrastructure. Appl Clin Inform. 2016;7:260–74.
45. Mamzer MF, Duchange N, Darquy S, et al. Partnering with patients in translational oncology research: ethical approach. J Transl Med. 2017;15:74.
46. Tabor HK, Stock J, Brazg T, et al. Informed consent for whole genome sequencing: a qualitative analysis of participant expectations and perceptions of risks, benefits, and harms. Am J Med Genet A. 2012;158A:1310–9.

47. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19:249–55.
48. McGowan ML, Settersten RA Jr, Juengst ET, Fishman JR. Integrating genomics into clinical oncology: ethical and social challenges from proponents of personalized medicine. Urol Oncol. 2014;32:187–92.
49. Savage N. Privacy: the myth of anonymity. Nature. 2016;537:S70–2.
50. Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name. Harvard University, Data Privacy Lab 1021-1: 2013.
51. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. Science. 2013;339:321–4.
52. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15:565–74.
53. Jarvik GP, Amendola LM, Berg JS, et al. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. Am J Hum Genet. 2014;94:818–26.
54. Wolf SM, Burke W, Koenig BA. Mapping the ethics of translational genomics: situating return of results and navigating the research-clinical divide. J Law Med Ethics. 2015;43:486–501.
55. Shalowitz DI, Miller FG. Communicating the results of clinical research to participants: attitudes, practices, and future directions. PLoS Med. 2008;5:e91.
56. Wolf SM, Annas GJ, Elias S. Point-counterpoint. Patient autonomy and incidental findings in clinical genomics. Science. 2013;340:1049–50.
57. Ross LF, Rothstein MA, Clayton EW. Mandatory extended searches in all genome sequencing: "incidental findings," patient autonomy, and shared decision making. JAMA. 2013;310:367–8.
58. Lolkema MP, Gadellaa-van Hooijdonk CG, Bredenoord AL, et al. Ethical, legal, and counseling challenges surrounding the return of genetic results in oncology. J Clin Oncol. 2013;31:1842–8.
59. Kronenthal C, Delaney SK, Christman MF. Broadening research consent in the era of genome-informed medicine. Genet Med. 2012;14:432–6.
60. Edwards KL, Korngiebel DM, Pfeifer L, et al. Participant views on consent in cancer genetics research: preparing for the precision medicine era. J Community Genet. 2016;7:133–43.
61. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008;9:406–11.
62. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15:409–21.
63. Beecher HK. The powerful placebo. J Am Med Assoc. 1955;159:1602–6.
64. Turner JA, Deyo RA, Loeser JD, et al. The importance of placebo effects in pain treatment and research. JAMA. 1994;271:1609–14.
65. Chvetzoff G, Tannock IF. Placebo effects in oncology. J Natl Cancer Inst. 2003;95:19–29.
66. Baldwin D, Broich K, Fritze J, et al. Placebo-controlled studies in depression: necessary, ethical and feasible. Eur Arch Psychiatry Clin Neurosci. 2003;253:22–8.
67. Chaput de Saintonge DM, Herxheimer A. Harnessing placebo effects in health care. Lancet. 1994;344:995–8.
68. Oliver RT. Are cytokine responses in renal cell cancer the product of placebo effect of treatment or true biotherapy? What trials are needed now? Br J Cancer. 1998;77:1318–20.
69. Shepherd FA, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med. 2005;353:123–32.
70. Burger RA, Brady MF, Bookman MA, et al. Incorporation of bevacizumab in the primary treatment of ovarian cancer. N Engl J Med. 2011;365:2473–83.
71. Zion D. Ethical considerations of clinical trials to prevent vertical transmission of HIV in developing countries. Nat Med. 1998;4:11–2.
72. Emanuel EJ. Reconsidering the Declaration of Helsinki. Lancet. 2013;381:1532–3.
73. Chiodo GT, Tolle SW, Bevan L. Placebo-controlled trials: good science or medical neglect? West J Med. 2000;172:271–3.
74. Mills EJ, Seely D, Rachlis B, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. Lancet Oncol. 2006;7:141–8.

75. Comis RL, Miller JD, Aldige CR, et al. Public attitudes toward participation in cancer clinical trials. J Clin Oncol. 2003;21:830–5.
76. Fede AB, Miranda MC, Lera AT et al. Placebo-controlled trials (PCT) in cancer research: patient and oncologist perspectives. In: 2010 ASCO annual meeting. Chicago. J Clin Oncol. 2010;e19626.
77. Miranda MC, Fede AB, Magalhaes N et al. Outcomes from placebo/best supportive care-controlled trials (PBSCT) in the era of molecular targeted therapy: a meta-analysis. In: 2010 ASCO annual meeting. Chicago: J Clin Oncol. 2010;6127.
78. Daugherty CK, Ratain MJ, Emanuel EJ, et al. Ethical, scientific, and regulatory perspectives regarding the use of placebos in cancer clinical trials. J Clin Oncol. 2008;26:1371–8.
79. Prasad V, Grady C. The misguided ethics of crossover trials. Contemp Clin Trials. 2014;37:167–9.
80. Gupta U, Verma M. Placebo in clinical trials. Perspect Clin Res. 2013;4:49–52.
81. Miranda Vda C, Fede AB, Lera AT, et al. How to consent without understanding? Rev Assoc Med Bras (1992). 2009;55:328–34.

Cost-Effectiveness Studies in Oncology 16

Pedro Aguiar Jr., Carmelia Maria Noia Barreto, Brittany L. Bychkovsky, and Gilberto de Lima Lopes Jr.

16.1 Background

Historically, cancer patients had few therapeutic options and universally poor outcomes. However, the field of oncology has advanced in the past decade, with better cancer screening protocols and new therapies, and survival outcomes for most cancer subtypes have improved in high-income countries [1]. While these advances should be applauded, they have come at a cost—a high cost that is not always affordable in many countries.

In high-income, middle-income, and low-income countries, cancer incidence is increasing, partly owing to aging populations, environmental factors (obesity, persistent tobacco use and exposure, etc.), and insufficient cancer prevention efforts; consequently cancer mortality is rising around the world [2]. In 2008, the economic impact of cancer due to premature mortality and morbidity was estimated to be US$895 billion worldwide [3]. In the 2000s, the global scientific community has responded to this pressing economic issue and there has been renewed interest (and investment) in cancer research and drug discovery. From 2010 to 2014, 25 new

P. Aguiar Jr., M.D.
Faculdade de Medicina do ABC, Santo André, SP, Brazil

C. M. N. Barreto, M.D.
Clinical Oncology Sector, Sociedade Beneficência Portuguesa
de São Paulo, São Paulo, SP, Brazil

B. L. Bychkovsky, M.D., M.Sc.
Dana-Farber Cancer Institute, Boston, MA, USA

Harvard Medical School, Boston, MA, USA
e-mail: Brittany_bychkovsky@dfci.havard.edu

G. de Lima Lopes Jr., M.D., M.B.A., F.A.M.S. (✉)
Global Oncology Program, Sylvester Comprehensive Cancer Center
at the University of Miami, Miami, FL, USA

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_16

drugs were approved for the treatment of cancer; an impressive figure totaling more than half the number of new drugs approved in the preceding four decades [4]. From 2010 to 2013 the number of new agents approved for the treatment of cancer superseded the value observed in the previous 10 years [4]. If this rate of growth continues, the 2010s may see up to 67 new cancer drugs enter the market [4]. To complicate matters, novel cancer drugs typically cost more and may be taken for longer periods of time than the older alternatives [4].

These trends represent the rising cost of cancer care. For example, data from the Brazilian Court of Auditors showed that the annual cost of treating cancer in Brazil had doubled between 2002 and 2008, from US$250 million to US$500 million [5]. The acceleration of cancer's economic burden is disproportionately high and is outpacing the gradual rise in median household income and inflation around the world [6]. Consequently, cancer care strains both global and regional health systems and, hypothetically, may contribute to system failures. Clinical practitioners and policy makers must address this issue in order to provide the optimal cancer care at an affordable cost.

16.2 Types of Pharmacoeconomic Studies

Four types of economic analysis are applied to health-care systems: cost-effectiveness analysis, cost-utility analysis, cost-minimization analysis, and cost-benefit analysis. Table 16.1 summarizes the main characteristics of each type of study.

16.2.1 Cost-Effectiveness Analysis

Cost-effectiveness analysis examines the incremental cost of an experimental drug over its incremental effectiveness compared with the control drug. The incremental cost is measured in monetary units. The effectiveness of each alternative is

Table 16.1 Types of pharmacoeconomic studies

Type of study	Cost	Benefit	Advantages	Disadvantages
Cost-effectiveness	Monetary units (drug costs)	Outcome units (e.g., survival time)	Simpler to perform	Do not consider quality of life
Cost-utility	Monetary units (direct and indirect costs)	Quality-adjusted life years (QALY)	Consider quality of life	Do not consider intangible costs
Q-TWiST	Monetary units (direct and indirect costs)	Quality-adjusted life years (QALY)	Consider before and after disease progression	Not useful for metastatic disease
Cost-minimization	Monetary units (direct and indirect costs)	None (treatments have the same proven efficacy)	Is the most reliable to compare generics and biosimilars	Drugs must have the same proven efficacy
Cost-benefit	Monetary units (direct and indirect costs)	Monetary units (intangible costs)	Is the gold standard for economic studies	Intangible costs are very difficult to assess

Q-TWiST Quality-adjusted time with and without symptoms or toxicity measured in natural units (life years gained, cases successfully treated, or cases averted). The main advantage of this type of study is that it is the most understandable type of pharmacoeconomic analysis. However, cost-effectiveness analysis considers only the acquisition costs of interventions and looks only at survival, but not at toxicity, inconvenience, or effects on quality of life (QoL). For example, surgery versus radiation for the primary treatment of a specific cancer can be compared by their costs per life year gained, but such a comparison may lead to an incorrect conclusion if the treatment effects on QoL are very different. In other words, cost-effectiveness analysis can help in choosing among treatments with similar efficacy and toxicity profiles for a specific disease, but not in making choices across dissimilar treatments and conditions [7, 8].

16.2.2 Cost-Utility Analysis

Cost-utility analyses compare the incremental cost of a new drug over incremental utility. Utility is a measure of the value attributed to a health state, usually measured in quality-adjusted life years (QALY). The main difference between cost-utility analysis and cost-effectiveness analysis is that cost-utility associates mortality and morbidity data into a single multidimensional measure, QALY [9]. QALY is a measure of the quantity of life gained by a treatment, weighted by the quality of that life. QALY is not disease-specific, so it allows comparisons of the relative efficiency of health-care interventions for different conditions. Each of these metric analytic techniques has its place. While cost-utility analyses are indicated for comparing toxic treatment options, a cost-effectiveness analysis may be better for deciding between two diagnostic strategies. Quality of life is reflected in utility (a measure of preference for a given health state, rated on a scale from 0—the worst imaginable health state, to 10—perfect health). QALY is the product of the average survival resulting from an intervention and the QoL provided by the treatment [10]. For example, Nafees et al. developed a study that found health state utilities for non-small cell lung cancer [11]. The utility value for stable disease with no toxicity was 0.653. If a new drug were to provide an additional survival of 9 months (0.75 years) compared with standard therapy, the incremental QALY would be 0.489 (QALY = 0.75 × 0.653). Nevertheless, QALY needs to be adjusted for adverse events. The same study by Nafees et al. estimated the value of disutility for several adverse events [11]. If the new drug cited above were to cause more febrile neutropenia (−0.09 per event) and nausea (−0.04 per event) compared with standard therapy, the incremental QALY would decrease according to the frequency of the adverse events.

The Health Utilities Index and the European Quality of Life (EuroQoL) instrument are tools that allow estimates of utilities. However, other QoL instruments used in clinical trials may not be fit to be converted into utilities.

Another strategy used to integrate the quality and quantity of life is the quality-adjusted time with and without symptoms or toxicity (Q-TWiST). This is especially useful when one looks at interventions that will have health effects persisting beyond the duration of treatment, such as those seen with adjuvant chemotherapy.

Q-TWiST is the sum of the quality adjusted (u) time spent undergoing treatment and experiencing toxicity of any grade (TOX), plus the time spent free of disease in perfect health (TWiST), plus the time spent experiencing symptoms in disease relapse (REL), also expressed as: Q-TWiST = $u_{\mathrm{TOX}} \times \mathrm{TOX} + u_{\mathrm{TWiST}} \times \mathrm{TWiST} + u_{\mathrm{REL}} \times \mathrm{REL}$ [12]. For example, a study by Jang et al. assessed the Q-TWiST for adjuvant chemotherapy for lung cancer [13]. Survival curves for the treatment and observation groups were partitioned into three health states: time with ≥grade 2 (early or late) chemotherapy-related toxicity (TOX), time in relapse (REL), and time without toxicity or relapse (TWiST). Then the authors calculated the Q-TWiST value according to the utility for each health state. In this study, adjuvant chemotherapy had a Q-TWiST gain of 6.7.

16.2.3 Cost-Minimization Analysis

Cost-minimization analysis compares strategies of proven equal effectiveness (such as generics or biosimilars) to determine which one is the least expensive. Resource utilization is the only significant difference between the options. The comparison is made between the direct costs of each intervention; the most affordable intervention is the winner. There is no assessment of treatment consequences. In oncology, cost-minimization studies are unusual, because cancer treatments hardly ever produce equivalent survival or QoL [14], and large randomized controlled trials that perform head-to-head comparisons of treatments with drugs of the same class are difficult to coordinate and fund, except for non-inferiority trials, which may be useful for price competition. Generally, these non-inferiority studies are performed with public funds, since these types of studies are not of interest to most pharmaceutical companies and industry players, except in the context of the development of biosimilars.

16.2.4 Cost-Benefit Analysis

Cost-benefit analyses give monetary value to the health benefits of an intervention. If the cost/benefit ratio is <1, the intervention is attractive. Cost-benefit analyses are, in theory, the gold standard of economic evaluation. QALYs are valued in monetary terms to obtain the absolute benefit of the intervention. These analyses always produce a monetary outcome, so different potential uses of resources can be compared. On the other hand, putting a monetary value on the often-intangible outcomes of health care, the value of a life, might be problematic [8].

16.3 Study Framework

There are some key methodological rules that cost-effectiveness studies must follow. The primary endpoint of such an economic study is the most cost-effective intervention. In the oncology scenario, a cost-utility analysis is mostly applied to choose the best intervention, considering that the toxicity of an anticancer therapy is as relevant as survival gain.

In the adjuvant treatment context, the use of Q-TWiST is preferable compared to cost-utility analysis. However, this metric should not be applied for dissimilar treatments and interventions. The results of a cost-minimization analysis may favor one intervention over another in a specific context. For example, a program of early discharge after major surgery might save hospital expenditure, but increase the cost of home care services, and this may offset the savings. For such reasons, critical appraisal of economic studies is very important for reaching a conclusion and guiding health policy.

16.4 Identification and Assessment of Costs and Benefits

The identification and assessment of all relevant costs and benefits will determine the quality of a cost-effectiveness study [8].

16.4.1 Costs

Authors should include in their study the following costs:

- Direct treatment costs: the costs of all oncology therapies used in the study. The authors may consider all treatment lines as direct treatment costs.
- Direct non-treatment costs: the costs of resources used for patients to gain access to and participate in treatment, such as transportation and accommodation.
- Indirect costs: the costs related to the treatment of all relevant adverse events, as well as the end-of-life care costs.
- Intangible costs: whenever the real-life data is available, the authors may include the estimated costs of depression, anxiety, or pain. They may convert all these costs into setting-dependent monetary costs in order to perform a cost-benefit analysis. However, physicians and policy makers may find that these intangible costs may not be generalizable to settings different from the one where the economic study was performed.

Prospective data collected as part of a clinical trial may be more complete and more accurate than retrospective data; as a result, prospective data are preferable, although a retrospective economic analysis could be cheaper and done more rapidly than a prospective study, especially when the analysis is not in a clinical trial context [8].

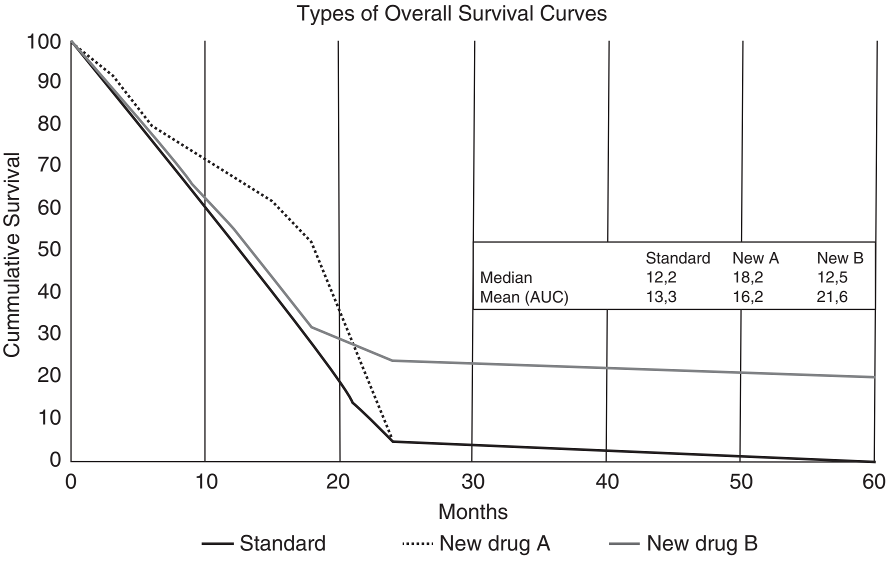
Importantly, all costs included in an economic analysis must be in the same context as those of any comparative study. For example, we do not recommend that the Brazilian Public Health System make policy decisions based on studies considering health-care costs in the United States.

16.4.2 Benefits

To analyze the benefit of an intervention/treatment in oncology, survival is the preferred outcome of interest. The survival data are often acquired from the most relevant randomized clinical trial available in the literature. A real-life study would be more reliable to demonstrate benefits, although such studies are expensive and not feasible for most pharmacoeconomic analyses.

A major concern is that sometimes a novel treatment demonstrates an impressive improvement in the median survival, while the tails of the Kaplan Meier probability curves of both the new and the standard interventions are similar. On the other hand, a novel therapy may not demonstrate any benefit in the median survival but may provide a significant difference in the long-term survival rate. In a cost-effectiveness study, because of these factors, it is important to determine the average benefit a patient can expect from a therapy, and this is made possible by determination of the area under the curve (AUC) for both of the treatments that are assessed in the study [8]. Figure 16.1 illustrates the importance of assessing the survival for the entire curve rather than assessing the median survival.

One of the most relevant concerns when assessing benefits is the time horizon of the analysis [15]. A majority of studies follow patients until the time at which it is expected that the main questions of the study will be answered. As a result, authors develop a model to estimate the survival benefit in a lifetime horizon. Sometimes, studies developed by manufacturers overestimate the survival benefit, and long-term follow-up analysis demonstrates that the real benefit is less robust than the benefit achieved by the survival model [15].



	Standard	New A	New B
Median	12,2	18,2	12,5
Mean (AUC)	13,3	16,2	21,6

Fig. 16.1 Examples of survival curves

It is also very important to consider all relevant adverse events when assessing treatment benefit. Generally, authors include literature-based values of disutility for each adverse event observed [11].

We should also discuss how the goals of cancer treatment are different for non-metastatic vs. metastatic disease. For patients without metastatic disease, we may accept treatments that have high rates of short-term toxicity if they allow for cure and have relatively low long-term toxicity. For patients with metastatic disease, it is all about QoL, and therefore our threshold for toxicities on therapy is much lower, because they negatively impact QoL.

16.5 Setting

For making any decision based upon a cost-effectiveness study, the setting of the analysis is very important. Economic evaluations are relatively specific to the health-care system in which they are performed [8].

Market forces, government regulations, and taxation law influence the costs of drugs. The differences among health-care systems worldwide make it difficult to translate the results of one economic study to a different context.

In terms of the benefits, when authors make a cost-utility analysis, they consider that the survival benefit must be adjusted to the QoL or utility provided by the treatment [9]. It is preferable to estimate the QoL provided by each treatment through a prospective analysis done in the same social context as the one that the cost-effectiveness study is designed to analyze [9]. Nevertheless, prospective QoL analyses can be very expensive and, to help investigators, several utility values have been published according to each disease-specific health status—although the accuracy of such QoL data may not be generalizable to every country or ethnicity [11]. Transparent and explicit disclaimers of the components of the analyses are crucial for proper study interpretation [8].

16.6 Transparency and Risk of Bias

Economic analyses evaluate many variables in order to reach a conclusion on whether or not a treatment is cost-effective. Some data may be unavailable and some data may be from trials with highly selected patients, and those trialists may, consciously or subconsciously, have ignored some data [8]. Because of these pitfalls, readers of cost-effectiveness studies must be especially critical of studies sponsored by pharmaceutical companies, as reports suggest that the industry often overestimates the benefits of treatments [15]. Therefore the authors of cost-effectiveness studies must disclose all information about the study with transparency, as well as disclosing any potential financial conflicts of interest associated with the study. This will help readers to make an impartial decision about an economic study.

16.7 Sensitivity Analysis

To analyze economic endpoints, large sample sizes are needed, even larger than those required to reach clinical endpoints, because there are variations in economic parameters (e.g. length of hospital stay). Therefore, detailed sensitivity analyses are necessary.

Sensitivity analyses are helpful, since they attempt to estimate resource use and the effectiveness of various interventions over a range of plausible possibilities (e.g. by using survival time confidence intervals). By varying the parameters input into the economic model, authors may depict what the study results would be if using assumptions different from those used in the base model. If the conclusion of an economic study changes with changes in the values of a key parameter, the specific parameter is considered sensitive, and the economic study is not robust [16].

Monte Carlo simulation allows us to assess all study parameters simultaneously. These analyses yield a cost-effectiveness acceptability curve, accounting for uncertainty in all model estimates. The result is that the curve displays the likelihood that a new intervention will have a cost-effectiveness ratio that falls below a particular societal “willingness to pay” [17].

Figure 16.2 illustrates the result of a Monte Carlo simulation. On the *X*-axis are the incremental values that payers or policy makers must consider to expend for each patient. This is the definition of “willingness to pay”. On the *Y*-axis are the probabilities of each drug being cost-effective for the incremental value being considered. Figure 16.2 shows that, with an additional investment of 100,000 dollars per patient, the probability of a new drug being cost-effective is 0%. On the other hand, with an additional 200,000 dollars invested per patient, the probabilities of new drug A and new drug B being cost-effective rise to 46% and 11%, respectively. The bold line in Fig. 16.2 is the “willingness to pay” threshold of the study, and this value is not fixed. The next section will discuss threshold issues.

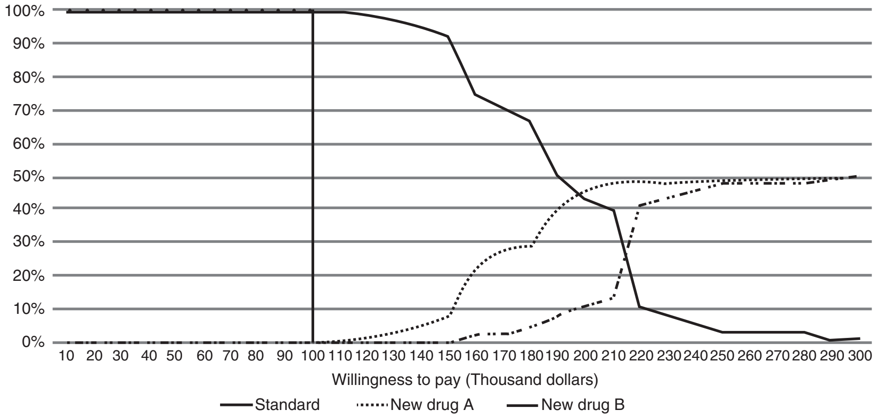


Fig. 16.2 Probability of being cost-effective

16.8 Cost-Effectiveness Thresholds

Cost-effectiveness thresholds are important for public health systems to preemptively define and use as a guide for whether the system will fund an intervention/service in health care. Most countries define cost-effectiveness thresholds as an amount per QALY [18]. For example, in Canada, an acceptable threshold is $20,000 per QALY for one intervention to be considered effective. In the United States, this value is $50,000 per QALY (the cost of 1 year of a dialysis treatment). In the United Kingdom, the maximum value accepted is £30,000 per QALY. Cost-effectiveness thresholds are sometimes arbitrary and may vary depending on each health-care scenario [19].

The World Health Organization considers therapies with an incremental cost for one additional QALY of less than three times the gross domestic product (GDP) per capita to be cost-effective, and those with a QALY of less than one GDP per capita are considered to be very cost-effective [17, 20]. It is very difficult to translate a health benefit into a value, and decisions on whether a certain intervention will be funded (or not funded) must not only consider the cost-effectiveness of the intervention, but also societal and community values. For example, it may be more appropriate to perform a bone marrow transplant in a child with leukemia (a rare occurrence, but an expensive intervention) than to offer an inexpensive intervention (say, dental care) at a population level, if saving the life of a child is more valuable for the community and society. Likewise, the cost of treating cancer must also be weighed against the cost of treating other chronic medical conditions [21].

Conclusion

Given that resources are finite, economic studies in oncology are important for guiding health policy decisions and the allocation of health-care resources in both the public and private sectors. Currently, cancer care has become exorbitantly expensive across the spectrum of high-, middle-, and low-income countries. The cost of cancer care is rising out of proportion to individual income, inflation, and the burdens of financing health-care systems around the world. In this setting, there is a need to emphasize health economic studies in oncology—more studies in this field are necessary, particularly in low- and middle-income countries. To summarize, in this chapter, we have discussed how and when to use different health economics studies and how such studies can help policy makers and physicians to prioritize resources.

References

1. Howlader N, Noone A, Krapcho M, et al. Cancer statistics review, 1975–2013. National Cancer Institute. https://seer.cancer.gov/csr/1975_2013/. Published 2016. Accessed 10 July 2017.
2. IARC. Fact Sheets by population. http://globocan.iarc.fr/Pages/fact_sheets_population.aspx. Accessed 27 July 2016.

3. American Cancer Society. The global economic cost of cancer 2010:12. http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf. Accessed 12 Sept 2016.
4. Savage P, Mahmoud S. Development and economic trends in cancer therapeutic drugs: a 5-year update 2010–2014. Br J Cancer. 2015;112(6):1037–41. https://doi.org/10.1038/bjc.2015.56.
5. Brazilian Court of Auditors. National Policies for Cancer Care. 2011:132.
6. Ward E, Halpern M, Schrag N, et al. Association of insurance with cancer care utilization and outcomes. CA Cancer J Clin. 2008;58(1):9–31. https://doi.org/10.3322/CA.2007.0011.
7. Provenzale D, Lipscomb J. Cost-effectiveness: definitions and use in the gastroenterology literature. Am J Gastroenterol. 1996;91(8):1488–93. http://www.ncbi.nlm.nih.gov/pubmed/8759647. Accessed 21 Mar 2017
8. Earle CC, Coyle D, Evans WK. Cost-effectiveness analysis in oncology. Ann Oncol. 1998;9(5):475–82. https://doi.org/10.1023/A:1008292128615.
9. Gudex C, Kind P. The Qaly tool kit. New York: Centre for Health Economics, University of York; 1988.
10. Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care. 1989;5(4):559–75. http://www.ncbi.nlm.nih.gov/pubmed/2634630. Accessed 21 Mar 2017
11. Nafees B, Stafford M, Gavriel S, Bhalla S, Watkins J. Health state utilities for non small cell lung cancer. Health Qual Life Outcomes. 2008;6:84. https://doi.org/10.1186/1477-7525-6-84.
12. Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. J Clin Oncol. 1989;7(1):36–44. https://doi.org/10.1200/JCO.1989.7.1.36.
13. Jang RW, Le Maître A, Ding K, et al. Quality-adjusted time without symptoms or toxicity analysis of adjuvant chemotherapy in non-small-cell lung cancer: an analysis of the National Cancer Institute of Canada Clinical Trials Group JBR.10 trial. J Clin Oncol. 2009;27(26):4268–73. https://doi.org/10.1200/JCO.2008.20.5815.
14. Weltens C, Kesteloot K, Vandevelde G, Van den Bogaert W. Comparison of plastic and Orfit® masks for patient head fixation during radiotherapy: precision and costs. Int J Radiat Oncol. 1995;33(2):499–507. https://doi.org/10.1016/0360-3016(95)00178-2.
15. Prasad V, Jesús KD, Mailankody S. The high price of anticancer drugs: origins, implications, barriers, solutions. Nat Rev Clin Oncol. 2017. https://doi.org/10.1038/nrclinonc.2017.31.
16. Coyle D. Statistical analysis in pharmacoeconomic studies. A review of current issues and standards. PharmacoEconomics. 1996;9(6):506–16. http://www.ncbi.nlm.nih.gov/pubmed/10160478. Accessed 24 Mar 2017
17. Claxton K, Martin S, Soares M, et al. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. Health Technol Assess (Rockv). 2015;19(14):1–504. https://doi.org/10.3310/hta19140.
18. Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. CMAJ. 1992;146(4):473–81. http://www.ncbi.nlm.nih.gov/pubmed/1306034. Accessed 24 Mar 2017
19. Chambers JD, Neumann PJ, Buxton MJ. Does Medicare have an implicit cost-effectiveness threshold? Med Decis Mak. 2010;30(4):E14–27. https://doi.org/10.1177/0272989X10371134.
20. Neumann PJ, Cohen JT, Weinstein MC. Updating cost-effectiveness — the curious resilience of the $50,000-per-QALY threshold. N Engl J Med. 2014;371(9):796–7. https://doi.org/10.1056/NEJMp1405158.
21. Bae YHJ, Mullins CD. Do value thresholds for oncology drugs differ from nononcology drugs? J Manag Care Pharm. 2014;20(11):1086–92. https://doi.org/10.18553/jmcp.2014.20.11.1086.

How to Undertake Outcomes Research in Oncology

17

Monika K. Krzyzanowska and Melanie Powis

17.1 Introduction

While "outcomes research" is a commonly used term, there is no universally accepted definition or consensus on what this area of research entails. Often, *health services research*, *comparative effectiveness research,* and *outcomes research* are used interchangeably; however, while the terms are interconnected, each seeks to address slightly different research questions (Fig. 17.1) [1]. *Health services research* has generally been defined as descriptive research that is conducted on a population-based cohort or at the system level to address policy-related questions to inform the organization, funding, and/or delivery of health care [2]. In contrast, *comparative effectiveness research* utilizes observational data to compare the benefits and/or harms of two or more alternative treatments in a real-world, routine care setting [3] to fill in gaps in data from randomized trials, such as uptake, long-term complications, and resource utilization, or where such data is missing from randomized trials [4]. While the focus of the field of *outcomes research* has evolved considerably, it continues to emphasize the evaluation of endpoints to examine effectiveness and improve patient care in real-world settings [5].

Traditionally, outcomes research focused on developing reliable measures of the quality of care delivered, and evaluating, through observational research, the relationship between aspects of care and multi-dimensional outcomes, such as health status, quality of life, process measures, resource utilization, and costs [1, 6, 7]. More recently, outcomes research in oncology has evolved to include the evaluation of patient-centered outcomes, such as experience with care, and preferences and continuity of care, as well as including pragmatic evaluations of interventions aimed at improving quality of care. For the purposes of this chapter, we will discuss

M. K. Krzyzanowska, M.D., M.P.H., F.R.C.P.C. (✉) • M. Powis, M.Sc.
Division of Medical Oncology & Hematology,
Princess Margaret Cancer Centre, Toronto, ON, Canada
e-mail: monika.krzyzanowska@uhn.ca

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_17

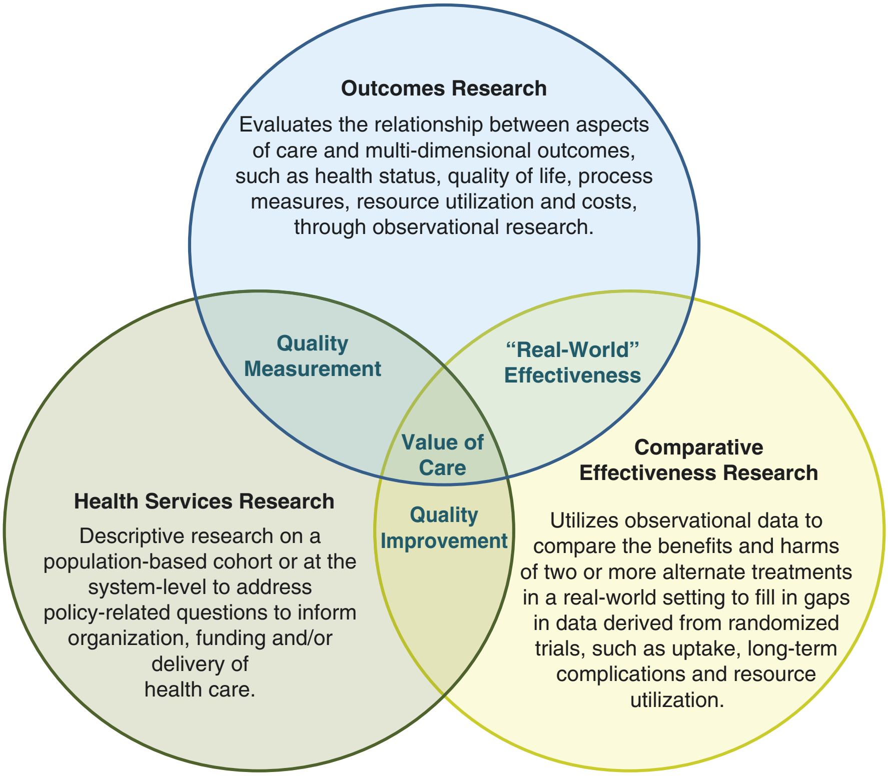

Outcomes Research
Evaluates the relationship between aspects of care and multi-dimensional outcomes, such as health status, quality of life, process measures, resource utilization and costs, through observational research.
Quality Measurement
"Real-World" Effectiveness
Value of Care
Quality Improvement
Health Services Research
Descriptive research on a population-based cohort or at the system-level to address policy-related questions to inform organization, funding and/or delivery of health care.
Comparative Effectiveness Research
Utilizes observational data to compare the benefits and harms of two or more alternate treatments in a real-world setting to fill in gaps in data derived from randomized trials, such as uptake, long-term complications and resource utilization.

Fig. 17.1 Outcomes research and related concepts

methods for measuring and improving the quality of cancer care, as well as observational and prospective, pragmatic evaluations of alternative treatments. We will close with a brief discussion of frameworks for evaluating value of care.

17.2 Historical Perspective

The origins of outcomes research can be traced to the early 1900s, when, as early as 1914, Codman argued that evaluation of the outcomes of care provided by a hospital was essential to determine whether the intended benefit was conferred by the treatment patients received [8]. He suggested the need for systematic reporting of past experiences, also suggesting that hospitals should not only report the number of patients treated, but instead be required to produce reports describing results of treatments delivered so that comparison with other institutions would be possible.

In 1966, Donabedian advanced Codman's ideas by coining the term *outcomes*, which he defined as measurable validators of effectiveness and quality of medicine [9]. In addition, he proposed the tripartite paradigm of structure-process-outcome as a systematic approach to assessing quality of care. In Donabedian's framework, "structure" refers to the physical and human capital resources available to deliver care, "process" refers to the actions/components of care delivery, and "outcomes" captures the effects of the care delivered or received on patients or populations, such as survival, restoration of function, or recovery. He called for the use of "quality measures" to systematically evaluate quality of care. A quality measure is a mathematical construct that usually consists of a numerator and denominator, whereby the denominator reflects the population eligible for a certain practice and the numerator captures the number of patients or institutions that received that practice.

In the 1970s and 1980s, advances in medical technology resulted in a rapid rise in the cost of health-care delivery, but there was little evidence to support improved patient outcomes [10]. In addition, Wennberg and Gittelsohn [11] reported wide variations in the use of various procedures and resources, rates of hospitalization, and costs associated with treatment, leading to a greater demand for evaluation and accountability for the care provided. Following a number of studies focusing on the systematic evaluation of practice variations in the United States in radiation and surgical oncology [12, 13] during this time period, it was recognized that, in order to move from observational studies evaluating small-scale practice variations to economically and clinically meaningful improvements in practice, a new paradigm was needed.

In 1988, Ellwood [14] called for a technology of *outcomes management* to link treatments to outcomes, encompassing the implementation of evidence-based practice standards and guidelines, collection of disease and quality-of-life outcomes in a database, use of these databases for systematic evaluation, and dissemination of the findings. By the end of the 1980s, however, it became increasingly apparent that supplemental primary data, beyond what was available in administrative databases, was needed to effectively evaluate new and established interventions [15].

17.3 Recent Trends in Outcomes Research

In the late 1990s, outcomes research moved away from a surveillance function and toward defining "appropriate care" to better inform patient-, provider-, and system-level decision-making, and aid third-party payers in optimizing resource utilization [16]. This involved the development and implementation of evidence-based care guidelines, practice standardization through continuous measurement, and benchmarking to define quality care and drive practice improvement [17]. In addition, comparative effectiveness studies began to be undertaken to compare alternative treatment strategies in a real-world setting [4], to define best practices and fill in gaps in knowledge—resulting from clinical trials having highly selected participants that are not necessarily representative of the disease populations from which they come [18]—to monitor disparities in adverse event reporting, and to take into account the fact that not all treatments could be evaluated in randomized trials owing to feasibility or ethical concerns.

More recently, the costs of delivering cancer care have risen dramatically as patient volumes increase, innovative high-cost treatments such as targeted and immunotherapies become increasingly available [19], and patients experience better disease outcomes, leading to the increased utilization of survivorship resources [20]. In response, policy makers and third-party payers have prioritized the need for new models of cancer care delivery and reimbursement that emphasize high-value care. As such, outcomes research has moved beyond surveillance and reporting to include the implementation and evaluation of interventions that address gaps in the quality of cancer care [21, 22]. While contemporary outcomes research is increasingly nested within randomized trials, these studies differ from traditional trials in that they focus on the process and delivery of care, with the main outcome being effectiveness, and they take a pragmatic approach to evaluation [23].

17.4 Outcomes Research Themes and Questions

A number of different methodologies are employed in outcomes research, the choice of which depends on the research question. Common areas of inquiry explored in outcomes research include: appropriateness of care, effectiveness, timeliness, equity, patient-centeredness, and value [23]. Studies evaluating appropriateness of care examine which patients are receiving what treatments and/or procedures, and usually evaluate potential over- or under-use of treatments. While much initial outcomes research has focused on the underutilization of evidence-informed practices or interventions, there has been increasing interest in the over-utilization of care. In oncology specifically, both the American Society of Clinical Oncology (ASCO) [24] and the American Society of Hematology (ASH) [25] have participated in the Choosing Wisely Campaign, started by the American Board of Internal Medicine to identify areas where there is little evidence to support clinical practice, as a first step towards decreasing overutilization. Effectiveness studies seek to evaluate medical interventions to fill in gaps from clinical trial data and examine their use in real-world settings. These studies examine how interventions are taken up or applied by clinicians, whether patients are adherent, and whether there is benefit following more wide-scale adoption. Studies of timeliness focus on barriers to patients accessing specific treatments or services, and may also evaluate the impact of time to accessing care on outcomes such as survival [26]. Equity studies evaluate disparities in the delivery of care to determine whether non-clinical factors such as race, sex, gender, or socioeconomic status influence care. Studies of patient-centeredness examine patient experience with care and care preferences, and may include evaluations of quality of life or continuity of care.

The four main types of research questions addressed in outcomes research in oncology include: (1) How is care delivered? (2) What is the quality of care? (3) Is an intervention or exposure effective? and (4) Is the care considered high-value? Summarized below are common approaches utilized to evaluate these research questions, as well as an overview of their applications and limitations.

17.5 Methods for Undertaking Outcomes Research

17.5.1 How Is Care Delivered: Evaluating Patterns of Care Delivery

The most common approach to studying how care is delivered is the retrospective cohort study, using either chart review or linked administrative data, or a combination of the two. Less common approaches include case-control studies and patient surveys. Regardless of the source of data used for the study, starting with clear objectives for the analysis prior to the data collection and analysis is essential.

17.5.1.1 Retrospective Chart Reviews

For studies evaluating care in small populations or in a single institution, retrospective manual chart review or hospital databases may be used [27]. Review of chart data generally provides greater clinical information regarding care, thereby providing greater contextual understanding of reasons for trends or disparities in the care being delivered. This may allow researchers to better examine the appropriateness of the care delivered, which is often impossible within larger, population-based studies; however, there are a number of limitations of chart reviews. For example, there may be fragmentation of documentation when aspects of care that span multiple institutions or settings are evaluated. An inherent limitation of retrospective chart reviews is the potential lack of generalizability to other populations, practices, and jurisdictions. Likewise, small sample sizes can increase the likelihood of confounding, and selection or measurement bias may affect the power of the study and the significance of the findings. Reviewing charts can also be time-consuming and expensive. As such, there has been significant interest in using population-based health-care billing data to look at trends in care delivery.

17.5.1.2 Administrative Data Analyses

In oncology specifically, researchers often utilize cancer registries that are deterministically linked to administrative billing databases, using unique identifiers, to answer questions surrounding care received in the routine practice setting. Administrative data is ideal for answering questions related to patterns of care, outcomes such as overall survival, performance measurement, and cost effectiveness. Depending on the comprehensiveness of the data holdings, the large amounts of available data that are routinely collected for billing may allow researchers to examine disparities [28] stemming from differences in socioeconomic status, sex, race, or geographical location, as well as system-, provider-, and patient-level drivers of outcomes. While utilizing administrative data to examine outcomes from a population-perspective offers many advantages, undertaking these analyses requires significant knowledge of how the data is structured within the databases and what data is available [1].

Health-care billing data contains records that span multiple institutions, so it can provide a fuller picture of certain aspects of care, such as resource utilization, than a chart review. As administrative data analyses utilize existing datasets that are collected in routine practice, they present a cost-effective method of assessing population-level trends in care. Clinical trial populations tend to be younger, healthier, and have higher socioeconomic status than the general population, factors that can pose issues with the generalizability of trial findings [29, 30]. Thus, the analysis of administrative data presents an avenue to explore the real-world uptake of guidelines and interventions by providers in routine care, as well as patient- and disease-related outcomes to fill in gaps in knowledge from clinical trials. In addition, the use of administrative data alleviates the likelihood of selection bias within the sample. However, unlike chart review studies, administrative data often lacks the clinical details needed to contextualize findings and evaluate the appropriateness of care received by patients [1, 31]. Likewise, the impact of patient or provider preferences on patterns of care cannot be evaluated. Since the data is not collected or structured for research, significant work is usually required to prepare the data for use and to operationalize the required variables before analysis can be undertaken. The steps taken in a typical analysis using administrative data are presented in Fig. 17.2. The completeness of administrative datasets is dependent on the jurisdiction and pay structure in place.

The types of databases utilized in these analyses may be hospital-specific, third-party payer, or larger and more comprehensive government holdings of data that capture billable health-care events. Administrative datasets are generally more comprehensive in universal, single-payer health-care systems where records account for the vast majority of episodes of care [32]. However, even in these systems, certain aspects of care that fall under private insurance, or where coverage is restricted to smaller sub-populations, may not be well captured in administrative data. Understanding the limitations of the administrative data that one is working with is essential for this type of outcomes research.

In jurisdictions with a multi-payer health-care system, such as the United States, the available administrative databases often only account for a small subset of the population. These datasets tend to be limited by market share, designation of hospital, rural vs. urban populations, geographical region of the country, and socioeconomic and racial differences [1]. Findings may vary dramatically when one evaluates databases derived from privately insured patients versus government-funded health care, such as—in the United States—Medicaid, which over represents patients with inherent social disparities in access to or delivery of care as it skews towards a lower socioeconomic status and higher percentage of ethnic minority patients. Accordingly, the generalizability of findings to other jurisdictions, even within the same country, may be limited.

17.5.2 What Is the Quality of Care: Measuring and Improving Health-Care Quality

Quality measurement is critical to understanding the quality of care delivered and driving system improvement. In 1999, the Institute of Medicine (IOM) published a report that described an ideal cancer care system and issued recommendations

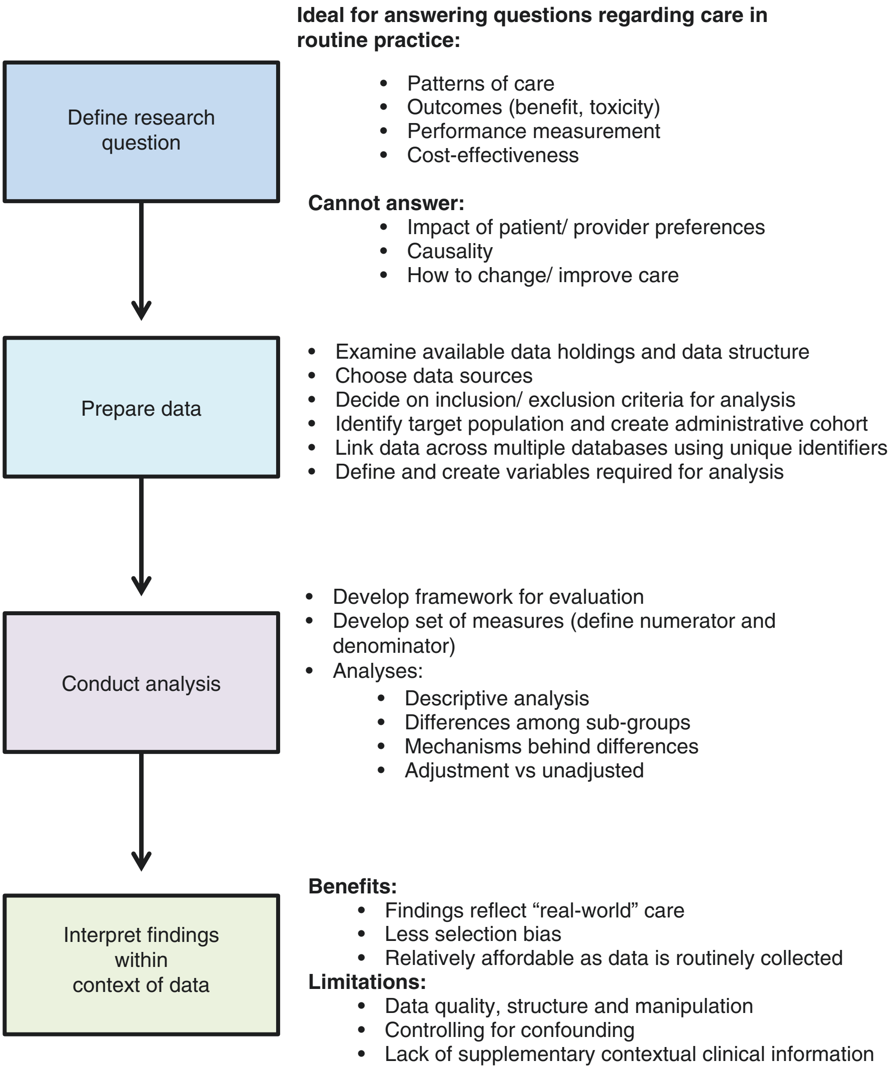

Define research question
Ideal for answering questions regarding care in routine practice:
• Patterns of care
• Outcomes (benefit, toxicity)
• Performance measurement
• Cost-effectiveness
Cannot answer:
• Impact of patient/ provider preferences
• Causality
• How to change/ improve care
Prepare data
• Examine available data holdings and data structure
• Choose data sources
• Decide on inclusion/ exclusion criteria for analysis
• Identify target population and create administrative cohort
• Link data across multiple databases using unique identifiers
• Define and create variables required for analysis
Conduct analysis
• Develop framework for evaluation
• Develop set of measures (define numerator and denominator)
• Analyses:
• Descriptive analysis
• Differences among sub-groups
• Mechanisms behind differences
• Adjustment vs unadjusted
Interpret findings within context of data
Benefits:
• Findings reflect "real-world" care
• Less selection bias
• Relatively affordable as data is routinely collected
Limitations:
• Data quality, structure and manipulation
• Controlling for confounding
• Lack of supplementary contextual clinical information

Fig. 17.2 Approach to using administrative data to evaluate the delivery of care

aimed at addressing gaps in the understanding and delivery of high-quality cancer care [33]. The report proposed six attributes of high-quality health care: that it be effective, safe, timely, efficient, equitable, and patient-centered, and recommended the development of a quality reporting program. Quality is usually assessed using quality measures or performance indicators, often as part of panels of measures. Substantial work has been done to determine the best approaches for defining quality measures and measuring quality. Determining the best use the findings to drive quality improvement and defining the best methods for quality improvement are areas that are less well developed and are now areas of active research.

17.5.2.1 Measuring Quality of Care

Quality measures are used to quantify, evaluate, and compare the quality of structures, processes, or outcomes of care being delivered [9]. Quality measures are mathematical constructs, consisting of a numerator and denominator, whereby the performance on a quality measure is usually expressed as the proportion of patients receiving a treatment or service relative to the number of patients that were eligible for that treatment or service. Defining and operationalizing quality measures allows researchers and decision-makers to evaluate quality of care, but deciding what to measure, and how, can be a challenge. Commonly cited criteria for selecting measures for either development or evaluation include the burden associated with the condition, the potential size and impact of the gap in care, the validity of the measure, and its actionability. To date, most quality measures have been process-based rather than structure or outcome measures, as measures of processes are considered to be more immediately actionable and easier to measure [34]. For example, a recent systematic review of studies that evaluated the quality of systemic therapy delivery in oncology found that access to treatment, which is an indicator of processes of care, was by far the most common domain of quality that has been examined [35].

The development of a panel of quality measures often begins with a review of the literature to identify existing indicators, followed by the prioritization of candidate measures, using consensus methods (Fig. 17.3). A commonly used consensus method is the modified Delphi panel process, which involves iterative rounds of ranking of candidate measures by a panel of experts to decide what measures to retain [36]. The candidate indicators are generally evaluated on validity, feasibility, and reliability, but the ranking criteria may vary depending on the purpose of the project. The need for the development of new indicators can also be assessed based on the availability and quality of existing indicators. When defining a novel quality measure, the concept to be measured, the target population, the risk adjustment strategy, the data sources, and the analysis plan must be established. These are called measure specifications.

Performance on a quality measure is usually reported as the proportion of patients, institutions, or providers meeting the criteria out of those eligible to receive or deliver the treatment or service. High variation in performance on a quality measure can highlight aspects of care with low standardization owing to insufficient evidence for informing best practice [37] or differences in the adoption of best practices. Quality measurement may be undertaken in small, local studies using retrospective chart review [38], or by utilizing large administrative datasets to evaluate and compare performance at the provider-, institution-, or system-level in order to inform policy [39]. The two main uses of quality measurement are for either quality improvement or for accountability. Regardless of the purpose of the measurement, understanding of the limitations of datasets used in quality measurement and understanding of how a measure was operationalized are essential for the appropriate interpretation of findings. Standardizing how a measure is defined can improve comparability and help improve interpretation.

One of the major limitations of panels of quality measures is that they are generally developed with a specific disease, phase of the cancer treatment continuum

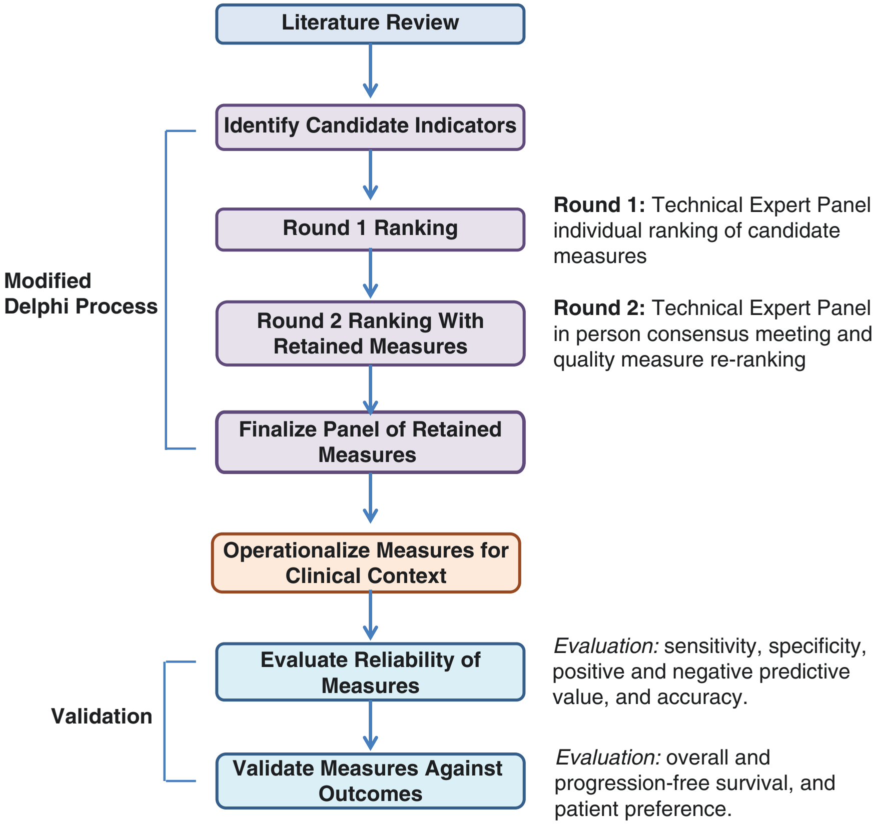

Literature Review
Identify Candidate Indicators
Round 1 Ranking
Round 2 Ranking With Retained Measures
Finalize Panel of Retained Measures
Modified Delphi Process
Round 1: Technical Expert Panel individual ranking of candidate measures
Round 2: Technical Expert Panel in person consensus meeting and quality measure re-ranking
Operationalize Measures for Clinical Context
Evaluate Reliability of Measures
Validate Measures Against Outcomes
Validation
Evaluation: sensitivity, specificity, positive and negative predictive value, and accuracy.
Evaluation: overall and progression-free survival, and patient preference.

Fig. 17.3 Process for developing a panel of quality indicators, using a modified Delphi method, and evaluation of their reliability and validity

(diagnosis, active treatment, or survivorship), or health system in mind, which can limit the generalizability and comparability of the findings. While administrative data allows institution-level performance trends to be observed, it does not take into account issues with access to health care or individual patient care preferences. The majority of quality measurement studies utilize administrative data, which lacks some of the contextual information required to assess the appropriateness of the care provided. As such, once a gap in care is identified using administrative data, additional evaluation may be needed to better understand how patient and provider factors influence performance, especially if quality improvement is planned.

An aspect of quality measurement that sometimes receives less attention is the issue of measure validation. Validation analyses complement the findings of quality measurement and are needed to understand how accurate and complete the data used to measure performance is and to evaluate the relevance of measures by examining the effects of performance on patient outcomes, such as survival and patient preference [40, 41]. For administrative data-derived measures, the reliability of the quality measure is often compared with data abstracted from patient charts. Concordance is then evaluated, and sensitivity, specificity, and negative and positive predictive values are calculated using the chart as the gold standard [42, 43]. These characteristics of individual measures can inform their interpretation and appropriate use. Univariate and multivariate regression models can be used to evaluate how adherence to quality measures relates to outcomes [40]; whether adjustment for patient or provider characteristics is needed to account for differences should also be considered.

17.5.2.2 Improving Quality of Care

The systematic measurement of health care quality can help to drive improvement through reporting [44] and provider incentives [45], but it can be challenging at times to define what "good quality" cancer care looks like and where to focus quality improvement efforts. Lack of standardization and unnecessary variation in inter-institutional and inter-provider practices in diagnosis, treatment, and surveillance have been identified as significant barriers to delivering high-quality cancer care [46]. There are a number of factors that may drive variation, including lack of consensus regarding best practices [47], resource availability, case-mix [48], and socioeconomic determinants of health [29]. It is important to understand which of these factors may be contributing to variation prior to the undertaking of improvement efforts, as these factors require different solutions.

Historically, targets for quality-improvement efforts have been selected based on their financial impact or perceived importance, rather than being based on using a systematic approach. Ideally, high-priority quality measures to target quality improvement should be meaningful, actionable, achievable, and have the potential to impact a large number of patients [47].

Targeting poor-performing quality measures with high variation for quality improvement work has been proposed as one methodology for ranking measures [49]. Hassett et al. [50] have proposed a prioritization framework that incorporates the degree of concordance, proportion of non-concordant patients in the population, and magnitude of clinical benefit in the ranking of quality measures for improvement. One limitation of this approach is the variation in relative clinical importance of quality measures over time, which may require re-ranking. An alternative methodology has been proposed by Enright et al. [51]; their methodology ranks measures based on the interquartile range of inter-institutional variation in performance and the eligible number of patients to generate a summative, priority rank. Once priorities for quality improvement have been identified, the next step is often to define performance targets.

17.5.2.3 Setting Performance Targets for Quality Improvement

Setting performance targets for quality measures can facilitate quality improvement by increasing adherence to evidence-based guidelines and improving patient outcomes; however, it can be difficult to determine what an appropriate target should be. Benchmarking can be used to set performance targets, identify top performers, and characterize factors associated with high performance on process indicators as a means to improve the quality of care being delivered [52]. Frequently, benchmarks are set subjectively, using consensus-derived approaches whereby aspirational performance targets are used to identify best practices and set standards of excellence; this method produces a framework for comparing performance, identifying the processes of top performers, and sharing best practices to improve quality of care [53, 54]. While this approach encourages quality improvement more effectively than audit and feedback alone, to be truly effective, performance targets must be realistic, attainable, and not unduly influenced by high performers with a low case volume [34]. Accordingly, target-setting activities have evolved to utilize data-driven approaches for deriving achievable targets [43]. Data-driven benchmarking generally utilizes the paired-mean approach developed by Kiefe et al. [34] to set objective performance targets by ranking institutions or providers in descending order of performance on a quality measure and calculating a target level of performance based on the proportion of patients in the top decile of the eligible population who meet the measure. Using this methodology, patients are assigned to a treating institution or provider, then these institutions or providers are ranked in descending order of performance on individual quality measures (Fig. 17.4). A subset cohort is generated by sequentially pooling patients until the combined size of the subset cohort is at least 10% of all eligible patients for the specific measure, starting from the

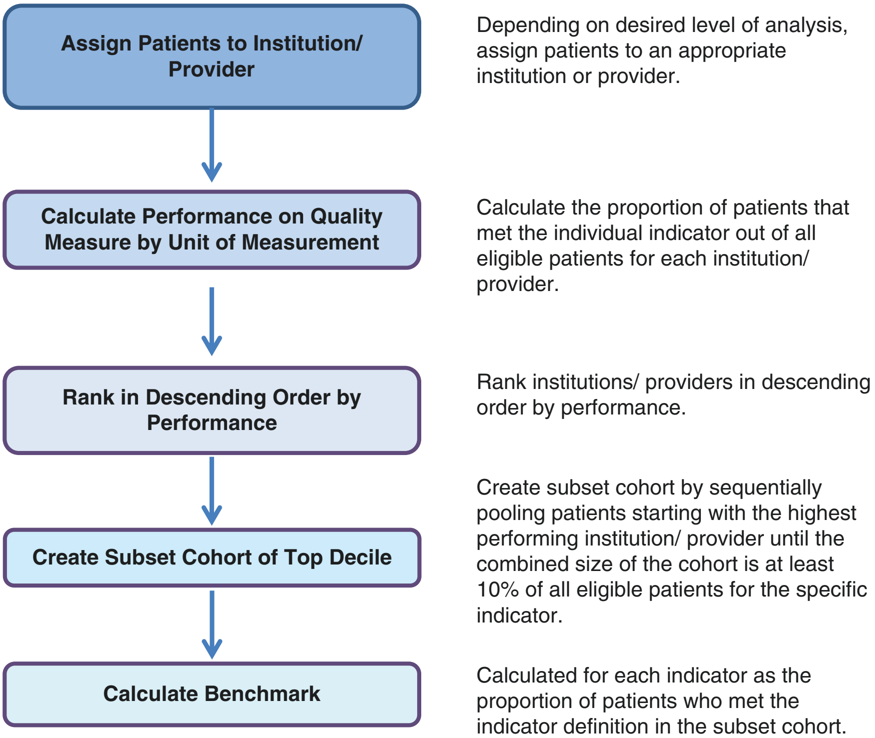

Assign Patients to Institution/ Provider
Depending on desired level of analysis, assign patients to an appropriate institution or provider.
Calculate Performance on Quality Measure by Unit of Measurement
Calculate the proportion of patients that met the individual indicator out of all eligible patients for each institution/ provider.
Rank in Descending Order by Performance
Rank institutions/ providers in descending order by performance.
Create Subset Cohort of Top Decile
Create subset cohort by sequentially pooling patients starting with the highest performing institution/ provider until the combined size of the cohort is at least 10% of all eligible patients for the specific indicator.
Calculate Benchmark
Calculated for each indicator as the proportion of patients who met the indicator definition in the subset cohort.

Fig. 17.4 Process for setting data-driven benchmarks for performance on quality measures using the methodology of Kiefe et al. [34]

highest performer. A benchmark performance rate, defined as the proportion of patients who met the indicator definition in this subset cohort, is then calculated for each indicator.

This methodology helps to avoid high-performing institutions or providers with a small case load unduly influencing the outcomes of the analysis, and is most suitable for evaluating process measures such as time to an event, or the use of a treatment or service [55, 56]. Benchmarking performance on outcome measures may require risk adjustment to account for the influences of patient-, provider-, or institution-level characteristics; however, an appropriate methodology has yet to be accepted [56]. Observing the dispersion of provider- or institution-level performance relative to the calculated benchmark can help to highlight indicators with higher inter-institutional variation that warrant further investigation into potential drivers of differences in performance. In these cases, further covariate analyses, evaluating institution-, provider-, and patient-level characteristics for individual indicators, may elucidate potential drivers of performance and help define the characteristics of high and low performers [57].

17.5.2.4 Frontline Quality Improvement

Methods used for health care quality improvement draw heavily on processes that have been utilized in the industrial sector since the early 1930s to drive system changes [58]. Continuous quality improvement focuses on system changes and treats every process as an opportunity for quality improvement. While the choice of quality improvement methodology depends on the nature of the project, the most commonly used method for rapid improvement utilizes cycles of "Plan-Do-Study-Act" (PDSA) [59]. This model systematically plans a change to a process that is expected to move the system closer to the desired state, implements the change, studies the effect of the change and deviations or defects that have occurred, and acts on the findings through additional iterative cycles of improvement [60]. In this model, run charts are employed; these are line graphs that are used to evaluate and visualize trends, patterns, and variation in data in response to process improvement efforts over time [61]. The vertical axis in the graph is the measure under study and the horizontal axis usually represents time, in units of days, weeks, months, or quarters, but alternative linear events may also be used, such as sequential patients, visits, or procedures. The median is plotted as a horizontal line with half of the observations above and half below the line. Run, trend, and shift tests are used to test for randomness in observed improvements or degradations in care. Additionally, horizontal lines representing performance targets can be added to the plots.

17.5.3 Is an Intervention or Exposure Effective: Observational and Prospective, Pragmatic Evaluations

The evaluation of new treatments and procedures in randomized controlled clinical trials is considered to be the gold standard for examining efficacy. However, clinical trial populations tend to be younger [29], have fewer comorbidities [30], and have higher socioeconomic status [62] than the general population, factors which can lead to issues with the generalizability of findings. Comparative effectiveness studies use observational data to evaluate interventions or exposures in the real-world setting. In recent years, there has also been growing interest in using prospective, but pragmatic, methods to evaluate interventions and models of care delivery in the routine care setting.

17.5.3.1 Comparative Effectiveness Studies

The goal of comparative effectiveness research is to compare the benefits and/or harms of two or more treatments in a real-world, routine care setting [3, 4]. Often these are retrospective observational studies that utilize administrative data to fill in gaps in knowledge from clinical trials by comparing the use of alternative treatment options in the routine care setting [5]. Additionally, these studies can examine how provider or institution characteristics affect the quality of a treatment [41]; evaluate clinical scenarios that are difficult to study in the trial setting because of low prevalence [63] or for ethical reasons; and provide estimates of other endpoints, such as long-term complications associated with treatment, costs, and resource utilization [64, 65]. While comparative effectiveness studies allow for the evaluation of how treatments and services are put into practice in routine care, such studies can still fall short of providing a full picture of the effectiveness of an intervention based on patient outcomes, such as quality of life, continuity of care or preferences, that are not traditionally collected in routine care and require the acquisition of supplementary primary data [66].

17.5.3.2 Prospective, Pragmatic Evaluations

Traditionally, outcomes research has evaluated the effects of different aspects of care on outcomes related to disease and process, utilizing observational data, but as noted above, observational data generally provides limited information on how different aspects of care affect patient-centered outcomes such as patient experience or continuity of care. These variables are not routinely captured, but can be as important as disease outcomes when evaluating care and examining the uptake of an intervention [67]. Recently, health-care priorities have included the evaluation and implementation of new models of patient-centered care to provide higher value care and address priority areas identified for quality improvement initiatives [68]. This, in turn, has led to the increasing integration of patient-centered outcomes into research and practice, and the need for prospective, pragmatic evaluations of interventions. For example, a number of these initiatives in oncology address the high rates of emergency department visits and hospitalizations in patients undergoing active treatment, factors which have been noted in multiple jurisdictions [65, 69, 70]. These initiatives have led to the development of new models of care, such as patient-centered medical homes and web-based interventions, to better manage cancer treatment-related symptoms that are being evaluated prospectively, using combinations of traditional disease outcomes, patient-centered outcomes, and health system metrics [71, 72].

17.5.4 Is It Considered High-Value Care: Defining Value in Cancer Care

Owing to the rising incidence of cancer and the increased costs of the delivery of innovative treatments [73], the concept of value has emerged as a dynamic and necessary component for evaluating novel interventions to ensure the best use of limited resources [74]. Current value frameworks focus on the active-treatment phase of cancer care [75, 76], and are largely based on cost-effectiveness analyses, which evaluate interventions based on rates of treatment-related toxicity and overall survival, usually measured in quality-adjusted life years, disability-adjusted life years, or incremental cost-effectiveness ratios [75, 77]. Value analyses seek to link the costs of different cancer treatments to their impact at the patient and population levels, in terms of benefits (clinical efficacy and cost) and toxicities, within the contexts of medical need, disease prevalence, and available alternative treatments [74, 75]. To examine value, numerous outcomes research methodologies, such as administrative data analyses coupled with reported outcomes for randomized controlled trials, are linked in order to calculate the composite value scores.

17.5.4.1 Value Frameworks

Value frameworks have evolved to provide standardization and transparency in health-care decision-making [77]. However, value is not a static concept; value is linked to context and can change depending on the jurisdiction, type of health-care system, disease site, treatment intent, and patient population under valuation. For example, quality of life might hold more value than overall survival in the advanced-care setting, so the weights of these attributes may be different from those in models that examine curative or adjuvant treatments. In addition, the patient perspective, which is central to the definition of value, is heavily individualized. As such, the value proposition from the patient's perspective considers not only clinical efficacy, but also quality of life and convenience, and is dependent on variables such as age, comorbidities, personal finances, and individual beliefs [75]. However, while these attributes can affect the uptake of practice recommendations, patient-centered outcomes can be difficult to quantify and are not routinely collected, so they are often omitted from current value frameworks. Thus, further work is needed to elucidate how best to incorporate patient perceptions of care into value frameworks.

17.6 Future Directions

Data collection within health-care systems is evolving rapidly with the greater adoption and uptake of electronic medical records in recent years, allowing for easier aggregation of data across providers, institutions, and health-care systems, although challenges with converting this data into usable formats exist and need to be addressed [78]. This greater access to data that can be used to fill in gaps in evidence when evaluating interventions and outcomes represents a crucial opportunity for outcomes researchers and policymakers in their pursuit of improved health care and lower costs. In addition, advances in "big data" in health informatics have the potential to allow for more timely receipt of patient-, provider-, and system-level data, which, in turn, could facilitate the measurement and improvement of the quality of care provided [79]. As a result, the concept of "learning healthcare systems" has recently emerged, whereby information is taken up in real time to improve care, or to provide decision support that is more generalizable to the population [80].

17.7 Summary

Outcomes research can be defined more by the types of questions it addresses than simply by the methods used to answer those questions. Outcomes research has evolved from its roots in measuring and reporting on aspects of cancer care, delivered using retrospective chart review and administrative data, to driving improvements in the quality of cancer care by using data for performance management and for filling in gaps in medical evidence generated by clinical trials. The next phase of outcomes research will continue to build upon these uses by taking advantage of big data and mixed-methods prospective evaluations of interventions to address issues in care delivery in oncology.

List of Technical Terms and Abbreviations

Administrative data Data collected routinely for billing purposes.

ASCO American Society of Clinical Oncology.

ASH American Society of Hematology.

Case-control study A type of observational study that compares a group with an existing outcome ("cases") with a similar group without the outcome ("controls") with respect to an exposure.

Cohort study A study in which researchers compare what happens to a group that has been exposed to a particular variable with a group who have not been exposed.

Comparative effectiveness research Utilizes observational data to compare the benefits and harms of two or more alternative treatments in a real-world, routine care setting to fill in gaps in data derived from randomized trials, such as uptake, long-term complications, and resource utilization, or where such data is missing from randomized trials.

Delphi panel A structured, systematic consensus process that utilizes iterative rounds of evaluation, performed by a panel of experts, to converge on an answer.

Health services research Descriptive research that is conducted on a population-based cohort or at the system level to address policy-related questions to inform the organization, funding, and/or the delivery of health care.

Outcomes research Emphasizes the use of endpoints to examine effectiveness and improve patient care.

PDSA Plan-Do-Study-Act; iterative cycles of quality improvement.

process measure Measures that evaluate actions or components of care delivery.

Quality indicators/measures Mathematical constructs, consisting of a numerator and denominator, that are used to quantify, evaluate, and compare the quality of structures, processes, or outcomes of care being delivered; usually expressed as the proportion of patients receiving a service relative to the number who were eligible to receive that service.

Run chart Simple line graph of a measure plotted against time; used to evaluate and visualize trends, patterns, and variation in data in response to process improvement efforts.

structure measure This indicator measures the physical and human capital resources available to deliver care.

References

1. In H, Rosen JE. Primer on outcomes research. J Surg Oncol. 2014;110(5):489–93.
2. Lohr KN, Steinwachs DM. Health services research: an evolving definition of the field. Health Serv Res. 2002;37(1):7–9.
3. Marko NF, Weil RJ. The role of observational investigations in comparative effectiveness research. Value Health. 2010;13(8):989–97.
4. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. J Clin Oncol. 2012;30(34):4215–22.
5. Wilkin D. Outcomes research in primary health care. Fam Pract. 1985;2(4):253–4.
6. Lipscomb J, Donaldson MS. Outcomes research at the National Cancer Institute: measuring, understanding, and improving the outcomes of cancer care. Clin Ther. 2003;25(2):699–712.
7. Karakiewicz PI, Briganti A, Chun FK, Valiquette L. Outcomes research: a methodologic review. Eur Urol. 2006;50(2):218–24.
8. Codman EA. The product of a hospital. Surg Gynecol Obstet. 1914;18:491–6.
9. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q. 1966;44:166–206.
10. Cochrane AL. Archie Cochrane in his own words. Selections arranged from his 1972 introduction to “Effectiveness and efficiency: random reflections on health services” 1972. Control Clin Trials. 1989;10:428–33.
11. Wennberg J, Gittelsohn A. Small area variations in health care delivery. Science. 1973;182:1102–8.
12. Hanks GE, Kramer S, Diamond JJ, Herring DF. Patterns of care outcome survey: national outcome data for six disease sites. Am J Clin Oncol. 1982;5:349–53.
13. Kramer S, Hanks GE, Diamond JJ, MacLean CJ. The study of the patterns of clinical care in radiation therapy in the United States. CA Cancer J Clin. 1984;34:75–85.
14. Ellwood PM. Shattuck lecture- outcomes management: a technology of patient experience. N Engl J Med. 1988;318:1549–56.

15. Institute of Medicine (US) Committee on Potential Conflicts of Interest in Patient Outcomes Research Teams. In: Donaldson MS, Capron AM, editors. Patient outcomes research teams: managing conflict of interest. Washington (DC): National Academies Press (US); 1991.
16. Epstein AM. The outcomes movement- will it get us where we want to go? N Engl J Med. 1990;323:266–7.
17. van Dam PA, Tomatis M, Marotti L, Heil J, Wilson R, Rosselli Del Turco M, et al. The effect of EUSOMA certification on quality of breast cancer care. Eur J Surg Oncol. 2015;41(10):1423–9.
18. Prince RM, Atenafu EG, Krzyzanowska MK. Hospitalizations during systemic therapy for metastatic lung cancer: a systematic review of real world vs clinical trial outcomes. JAMA Oncol. 2015;1(9):1333–9.
19. Fojo T, Mailankody S, Lo A. Unintended consequences of expensive cancer therapeutics—the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley lecture. JAMA Otolaryngol Head Neck Surg. 2014;140:1225–36.
20. Parry C, Kent EE, Mariotto AB, Alfano CM, Rowland JH. Cancer survivors: a booming population. Cancer Epidemiol Biomark Prev. 2011;20(10):1996–2005.
21. Mooney K, Beck SL, Wong B, Dunson WA, Wujcik D. An IT-integrated, computer-based telephone system for monitoring patient-reported symptoms: result of two trials. J Clin Oncol. 2013;30(suppl 34):abst 2.
22. Kearney N, McCann L, Norrie J. Evaluation of a mobile phone-based, advanced symptom management system (ASyMS) in the management of chemotherapy-related toxicity. Support Care Cancer. 2009;17(4):437–44.
23. Lee SJ, Earle CC, Weeks JC. Outcomes research in oncology: history, conceptual framework, and trends in the literature. J Natl Cancer Inst. 2000;92(3):196–204.
24. Schnipper LE, Lyman GH, Blayney DW. American Society of Clinical Oncology 2013 top five list in oncology. J Clin Oncol. 2013;31:4362–70.
25. Hicks LK, Bering H, Carson KR, Kleinerman J, Kukreti V, Ma A, et al. The ASH cChoosing Wisely® campaign: five hematologic tests and treatments to question. Blood. 2013;122(24):3879–83.
26. Biagi JJ, Raphael MJ, Mackillop WJ, Kong W, King WD, Booth CM. Association between time to initiation of adjuvant chemotherapy and survival in colorectal cancer: a systematic review and meta-analysis. JAMA. 2011;305(22):2335–42.
27. Juarez JE, Choi J, St John M, Abemayor E, TenNapel M, Chen AM. Patterns of care for elderly patients with locally advanced head and neck cancer. Int J Radiat Oncol Biol Phys. 2017;98:767–74. https://doi.org/10.1016/j.ijrobp.2017.01.209. pii: S0360-3016(17)30271-7 [Epub ahead of print]
28. Fountzilas C, Chang K, Hernandez B, Michalek J, Crownover R, Floyd J, Mahalingam D. Clinical characteristics and treatment outcomes of patients with colorectal cancer who develop brain metastasis: a single institution experience. J Gastrointest Oncol. 2017;8(1):55–63.
29. Murthy VH, Krumholtz HM, Gross CP. Participating in cancer clinical trials: race-, sex-, and age-based disparities. JAMA. 2004;291:2720–6.
30. Simon MS, Du W, Flaherty L, Philip PA, Lorusso P, Miree C, et al. Factors associated with breast cancer clinical trials participation and enrollment at a large academic medical centre. J Clin Oncol. 2004;22:2046–52.
31. Enright KA, Krzyzanowska MK. Benefits and pitfalls of using administrative data to study hospitalization patterns in patients with cancer treated with chemotherapy. J Oncol Pract. 2016;12(2):140–1.
32. Williams J, Young W. A summary of studies on the quality of healthcare administrative databases in Canada. In: Goel V, Williams JI, Anderson GM, Blackstien-Hirsch P, Fooks C, Naylor CD, editors. Patterns of healthcare in Ontario: the ICES practice atlas. Ottawa: Canadian Medical Association; 1996. p. 339–45.

33. Institute of Medicine. Ensuring Quality of Cancer Care. The National Academies of Science and Engineering. https://www.nap.edu/catalog/6467/ensuring-quality-cancer-care. Accessed 22 April 2017.
34. Kiefe CI, Weissman NW, Allison JJ, Farmer R, Weaver M, Williams OD. Identifying achievable benchmarks of care: concepts and methodology. Int J Qual Health Care. 1998;10:443–7.
35. Shen S, Krzyzanowska MK. A decade of research on the quality of systemic cancer therapy in routine care: what aspects of quality are we measuring? J Oncol Pract. 2015;11(1):55–61.
36. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Information & Management. 2004;42(1):15–29.
37. In H, Neville BA, Lipsitz SR, Corso KA, Weeks JC, Greenberg CC. The role of National Cancer Institute-designated cancer center status: observed variation in surgical care depends on the level of evidence. Ann Surg. 2012;255(5):890–5.
38. American Society of Clinical Oncology: quality oncology practice initiative. http://www.instituteforquality.org/quality-oncology-practice-initiative-qopi. Accessed 16 May 2017.
39. Duvalko KM, Sherar M, Sawka C. Creating a system for performance improvement in cancer care: Cancer Care Ontario's clinical governance framework. Cancer Control. 2009;16(4):293–302.
40. Hassett MJ, Neville BA, Weeks JC. The relationship between quality, spending and outcomes among women with breast cancer. J Natl Cancer Inst. 2014;106(10):dju242.
41. Simunovic M, Rempel E, Thériault ME, Coates A, Whelan T, Holowaty E, et al. Influence of hospital characteristics on operative death and survival of patients after major cancer surgery in Ontario. Can J Surg. 2006;49(4):251–8.
42. Mandelblatt JS, Huang K, Makgoeng SB, Luta G, Song JX, Tallarico M, et al. Preliminary development and evaluation of an algorithm to identify breast cancer chemotherapy toxicities using electronic medical records and administrative data. J Oncol Pract. 2015;11(1):e1–8.
43. Earle CC, Neville BA, Landrum MB, Souza JM, Weeks JC, Block SD, et al. Evaluating claims-based indicators of the intensity of end-of-life cancer care. Int J Qual Health Care. 2005;17(6):505–9.
44. Hibbard JH, Stockard J, Tusler M. Hospital performance reports: impact on quality, market share, and reputation. Health Aff (Millwood). 2005;24(4):1150–60.
45. Grossbart SR. What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Med Care Res Rev. 2006;63(1 Suppl):29S–48S.
46. Aiello Bowles EJ, Tuzzio L, Wiese CJ, Kirlin B, Greene SM, Clauser SB, Wagner EH. Understanding high-quality cancer care: a summary of expert perspectives. Cancer. 2008;112(4):934–42.
47. Howard DH. A better understanding of variation in cancer care. Med Care. 2012;50(5):363–5.
48. Morgan J, Richards P, Ward S, Francis M, Lawrence G, Collins K, et al. Case-mix analysis and variation in rates of non-surgical treatment of older women with operable breast cancer. Br J Surg. 2015;102(9):1056–63.
49. Selby JV, Schmittdiel JA, Lee J, Fung V, Thomas S, Smider N, et al. Meaningful variation in performance: what does variation in quality tell us about improving quality? Med Care. 2010;48(2):133–9.
50. Hassett MJ, Hughes ME, Niland JC, Ottesen R, Edge SB, Bookman MA, et al. Selecting high priority quality measures for breast cancer quality improvement. Med Care. 2008;46(8):762–70.
51. Enright KA, Taback NA, Powis M, Gonzalez A, Yun L, Sutradhar R, et al. Setting quality improvement priorities for women receiving systemic therapy for early stage breast cancer using population level administrative data. J Clin Oncol. 2017;35:3207–14.
52. Thonon F, Watson J, Saghatchian M. Benchmarking facilities providing care: an international overview of initiatives. SAGE Open Med. 2015;3:2050312115601692.
53. van Dam PA, Verkinderen L, Hauspy J, Vermeulen P, Dirix L, Huizing M, et al. Benchmarking and audit of breast units improves quality of care. Facts Views Vis Obgyn. 2013;5(1):26–32.

54. Brucker SY, Schumacher C, Sohn C, Rezai M, Bamberg M, Wallwiener D. Steering Committee. Benchmarking the quality of breast cancer care in a nationwide voluntary system: the first five-year results (2003-2007) from Germany as a proof of concept. BMC Cancer. 2008;8:358.
55. Palmer RH. Process-based measures of quality: the need for detailed clinical data in large health care databases. Ann Intern Med. 1997;127(8 Pt 2):733–8.
56. Weissman NW, Allison JJ, Kiefe CI, Farmer RM, Weaver MT, Williams OD, et al. Achievable benchmarks of care: the ABCs of benchmarking. J Eval Clin Pract. 1999;5(3):269–81.
57. Powis M, Sutradhar R, Gonzalez A, Enright KA, Taback NA, Booth CM, et al. Establishing achievable benchmarks for quality improvement in systemic therapy for early-stage breast cancer. Cancer. 2017;123:3772–80.
58. Deming WE. Out of the crisis. Cambridge, MA: Center for Advanced Engineering Study; 1986.
59. Langley GJ, Nolan KM, Nolan TW, Norman CL, Provost LP. The improvement guide: a practical approach to enhancing organizational performance. San Francisco, CA: Jossey-Bass; 1996.
60. Plsek PE. Quality improvement methods in clinical medicine. Pediatrics. 1999;103(1 Suppl E):203–14.
61. Perla RJ, Provost LP, Murray SK. The run chart: a simple analytical tool for learning from variation in healthcare processes. BMJ Qual Saf. 2011;20(1):46–51.
62. Gross CP, Filardo G, Mayne ST, Krumholz HM. The impact of socioeconomic status and race on trial participation for older women with breast cancer. Cancer. 2005;103:483–91.
63. Crew KD, Neugut AI, Wang X, Jacobson JS, Grann VR, Raptis G, Hershman DL. Racial disparities in treatment and survival of male breast cancer. J Clin Oncol. 2007;25(9):1089–98.
64. Barzi A, Lenz HJ, Quinn DI, Sadeghi S. Comparative effectiveness of screening strategies for colorectal cancer. Cancer. 2017;123(9):1516–27.
65. Lee L, Crump M, Khor S, Hoch JS, Luo J, Bremner K, et al. Impact of rituximab on treatment outcomes of patients with diffuse large b-cell lymphoma: a population-based analysis. Br J Haematol. 2012;158(4):481–8.
66. Patel MI, Periyakoil VS, Blayney DW, Moore D, Nevedal A, Asch S, et al. Redesigning cancer care delivery: views from patients and caregivers. J Oncol Pract. 2017;13(4):e291–302.
67. Jensen RE, Snyder CF, Abernethy AP, Basch E, Potosky AL, Roberts AC, et al. Review of electronic patient-reported outcomes systems used in cancer clinical care. J Oncol Pract. 2014;10(4):e215–22.
68. Spandio JD. Oncology patient–centered medical home. J Oncol Pract. 2012;8(3 Suppl):47s–9s.
69. Enright K, Grunfeld E, Yun L, Moineddin R, Ghannam M, Dent S, et al. Population-based assessment of emergency room visits and hospitalizations among women receiving adjuvant chemotherapy for early breast cancer. J Oncol Pract. 2015;11(2):126–32.
70. Hassett MJ, O'Malley AJ, Pakes JR, Newhouse JP, Earle CC. Frequency and cost of chemotherapy-related serious adverse effects in a population sample of women with breast cancer. J Natl Cancer Inst. 2006;98(16):1108–17.
71. Kearney N, Miller M, Maguire R, Dolan S, MacDonald R, McLeod J, et al. WISECARE+: results of a European study of a nursing intervention for the management of chemotherapy-related symptoms. Eur J Oncol Nurs. 2008;12(5):443–8.
72. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol. 2016;34(6):557–65.
73. Committee on Improving the Quality of Cancer Care. Addressing the challenges of an aging population; Board on Health Care Services. In: Institute of Medicine, Levit L, Balogh E, Nass S, Ganz PA, editors. Delivering high-quality cancer care: charting a new course for a system in crisis. Washington (DC): National Academies Press (US); 2013.
74. Goulart BHL. Value: the next frontier in cancer care. Oncologist. 2016;21:651–3.
75. Schnipper LE, Davidson NE, Wollins DS, Tyne C, Blayney DW, Blum D, et al. American Society of Clinical Oncology Statement: a conceptual framework to assess the value of cancer treatment options. J Clin Oncol. 2015;33(23):2563–77.

76. Cherny NI, Sullivan R, Dafni U, Kerst JM, Sobrero A, Zielinski C, et al. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). Ann Oncol. 2015;26(8):1547–73.
77. Mandelblatt JS, Ramsey SD, Lieu TA, Phelps CE. Evaluating frameworks that provide value measures for health care interventions. Value Health. 2017;20(2):185–92.
78. Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. J Clin Oncol. 2012;30(34):4243–8.
79. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform. 2017;98:22–32.
80. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision. J Clin Oncol. 2016;34(7):756–63.

Systematic Reviews and Meta-Analyses of Oncology Studies

18

Allan A. Lima Pereira and Andre Deeke Sasse

18.1 Introduction

The current volume of research articles published every year is in continuous growth and it has become virtually impossible for physicians, even when they are focused on more specific fields, to keep up with the enormous amount of research data. The major reason for conducting a review is that large quantities of information must be simplified into palatable parts for understanding.

There are different types of reviews. Not all review articles are systematic reviews and not all systematic reviews are followed by a meta-analysis. Reviews that do not use planned scientific methods to search, collect, and summarize information are not systematic reviews. They usually are traditional narrative reviews, where there are no clearly specified methods of identifying, selecting, and validating information included from multiple studies. Once systematic reviews have been performed, only a subset of them will include statistical methods to quantify and combine the results from independent studies, which we call meta-analysis.

Commonly in oncology, there are controversies about the real value of interventions. It is, therefore, important to recognize potential biases and also to establish as accurately as possible the actual differences between the strategies being evaluated. Summarizing the evidence facilitates the interpretation of the results, and makes it possible to identify whether the claimed statistically significant benefits are also

A. A. Lima Pereira, M.D., Ph.D.
Department of Gastrointestinal Medical Oncology,
University of Texas – M.D. Anderson Cancer Center, Houston, TX, USA

A. D. Sasse, M.D., Ph.D. (✉)
Department of Internal Medicine, Faculty of Medical Sciences,
University of Campinas (UNICAMP), Campinas, SP, Brazil
e-mail: sasse@cevon.com.br

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_18

clinically relevant. For this reason, systematic reviews are needed to refine the unmanageable amounts of information found in electronic databases, separating the insignificant, unsound, or redundant deadwood in the literature from the studies that are worthy of reflection [1], and then using the processed information for different purposes, such as to:

- Make recommendations for clinical practice and guidelines
- Establish the state of existing knowledge (useful when applying for grants)
- Clarify conflicting data from different studies
- Highlight areas where further original research are required.

Also, many times, a meta-analysis can add better quality evidence to the current medical literature. For instance, after pooling together many underpowered negative studies, a meta-analysis can finally give us the answer that each study alone was unable to provide. However, if not done properly, a meta-analysis can lead to bias (metabias). In addition, systematic reviews have become impressively more common [2]. Therefore, it is crucial that physicians become familiar with interpreting this kind of work; the best way to do this is to gain understanding of the key points needed to perform such work.

18.2 How to Plan a Systematic Review

The first step in performing a systematic review is to define the research question. However, to avoid waste of time or duplication of efforts, it is important to search for published and ongoing systematic reviews which might have already answered the same question or are aiming to do so. This search can be made in specific databases, such as the Cochrane Library (http://www.cochranelibrary.com) and PROSPERO (www.crd.york.ac.uk/PROSPERO). General databases (e.g., MEDLINE and EMBASE) should also be searched.

After the research question has been decided and the need for a new review has been confirmed, a protocol should be registered in public databases (such as Cochrane and PROSPERO). A written protocol defines the study methodology and sets the inclusion/exclusion criteria for trials, literature searches, data extraction and management, assessment of the methodological quality of individual studies, and data synthesis. As for any clinical study, the systematic review protocol must be designed a priori. Although the majority of oncology medical journals do not require an a priori registered protocol, we believe this is necessary to minimize the risk of systematic errors or biases being introduced by decisions that are influenced by the findings.

Ideally, a systematic review and its protocol are planned and conducted by a team with multiple skills. A team leader should coordinate and write the final report. A medical oncologist with clinical practice is needed to clarify issues related to the chosen topic. Reviewers are required to screen abstracts, read the full text, and extract the data. A statistician can assist with data analysis. Frequently, researchers accumulate different functions, but a well-planned team helps to reduce the risk of errors; a team of at least three people is needed.

18.2.1 Framing the Question

As mentioned above, the beginning of a systematic review occurs through building a good clinical question. A well-formulated question usually has four parts: the population, the intervention; the comparison intervention; and the outcome. This question structure is known by the acronym PICO (Problem/Patient/Population, Intervention/Indicator, Comparison, Outcome). The PICO framework helps to identify key concepts of the question, and should be sufficiently broad to allow examination of variation in the study factor (e.g., intensity or dose regimen) and across populations. An example of a good and straightforward clinical question using the PICO framework can be found in a published systematic review [3] and is detailed below:

- P: metastatic colorectal cancer patients receiving first-line systemic palliative treatment
- I: complete stop of treatment
- C: continuous treatment until disease progression
- O: overall survival.

Therefore, the question is: “Does complete stop of treatment in the first-line palliative setting of metastatic colorectal cancer patients impact overall survival?” Note this final question allows the inclusion of different regimens, durations, and intensities, and makes it possible to evaluate only the strategy of concern. The decision of how broad or narrow a clinical question to use is based on clinical judgment. A “narrower” question may not be clinically useful and can result in false or biased conclusions. On the other hand, broad questions may pool together studies too different to be combined (“apples with oranges”) and make the search process more difficult and time-consuming.

Framing the question is not only the first step of a systematic review. It is also the most important, since it will have a direct impact on the inclusion and exclusion criteria used to select studies, the development of the search strategy, and the main data to be abstracted.

18.2.2 Searching the Evidence

It is easy to find a few relevant articles by a straightforward literature search, but the process becomes progressively more challenging as we try to find more “hidden” trials.

Systematic reviews of interventions require a thorough, objective, and reproducible search of a range of sources to identify as many relevant studies as possible. A search of PubMed/MEDLINE alone is not considered adequate. It is known that only 30% to 80% of all known published randomized trials are identifiable using MEDLINE [4]. In the field of oncology, it is critical to search electronic databases such as MEDLINE and EMBASE, but also databases from clinical trials, and summaries as the Cochrane Library. However, searching the LILACS database is irrelevant in systematic reviews in oncology [5].

It is essential to define in advance structured and highly sensitivity search strategies for the identification of trials in each database. These strategies should be described later in the formal article, to allow reproducibility. There are no magic formulae to make all of the process easy, but there are some standard tactics which could be helpful.

A central tactic for a good literature search in the electronic databases is to take a systematic approach to breaking down the review question into components, which can be combined using "AND" and "OR" terms. Using the example above, in the review evaluating "Does complete stop of treatment in the first-line palliative setting of metastatic colorectal cancer patients impact overall survival?", the key components:

- (colorectal neoplasms AND maintenance chemotherapy) represent the overlap between these two terms and retrieve only articles that use both terms. A PubMed search using these terms retrieved 279 articles (at the time of all searches, in April, 2017: new citations are added to the PubMed database regularly).
- (colorectal neoplasms AND (maintenance chemotherapy OR intermittent chemotherapy)) represents a broader search, which includes other possible terms in the articles that can describe the strategies. A PubMed search using these terms retrieved 513 articles.
- (colorectal neoplasms AND maintenance chemotherapy AND intermittent chemotherapy) represents the small set where all three terms overlap. A PubMed search using these terms retrieved only 13 articles.
- (colorectal neoplasms AND (maintenance chemotherapy OR intermittent chemotherapy) AND random*) combines the term random*, which is the shorthand for words beginning with random, e.g., randomized, randomization, randomly. A PubMed search using these terms retrieved 20 articles.

Although the overlap of all three terms will usually have the best concentration of relevant articles, this strategy will probably miss many relevant studies. The ideal search strategy combines precision with sensitivity.

Usually, the initial strategy will inevitably miss useful terms, and the search process will need to be repeated and refined. However, the results of initial searches are used to retrieve the initial relevant papers, which can be used in two ways to identify missed trials:

- The bibliographies of the found articles can be checked for articles missed by the initial search;
- A citation search, using the Science Citation Index, can be conducted to identify papers that have cited the identified studies, some of which may report subsequent primary research.

The missed paper can provide clues on how the search may be broadened to capture further papers, sometimes using other keywords. The whole process may then be repeated using the new keywords identified.

It is important to remember that studies are conducted in all parts of the world, and may be published in different languages. Ideally, a systematic review should include all relevant studies, irrespective of the publication language. Including articles written only in English would lead to greater biases, as positive studies conducted in countries where English is not the state language are more likely than negative ones to be submitted to an English-language journal. This increases the usual publication bias with an additional "tower of Babel" bias.

Having a reviewer who has good experience with databases is crucial for building an efficient literature search. But the use of multiple strategies is important to track down all relevant articles. As the whole process is complex and has a high risk of loss due to fatigue, it is fundamental that the literature searches should be done by two researchers, independently.

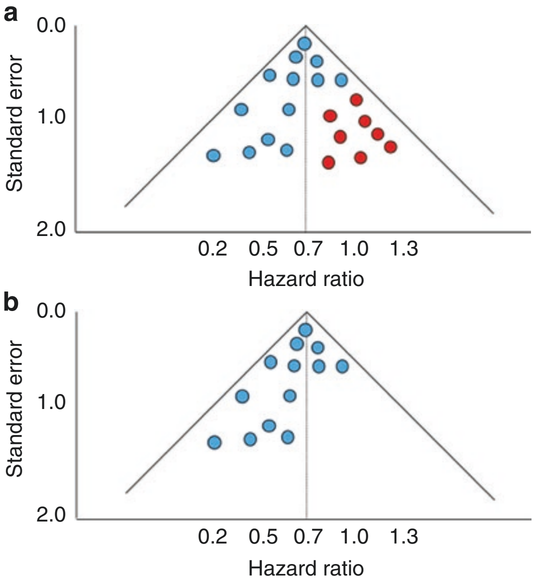
Duplicate publications and reports should be handled with caution. Systematic reviews have studies as the primary units of interest and analyses. However, a single study may have more than one report about the results. Each report should be analyzed and each may contribute useful information for the review. Thus, no publication should be discarded solely because of duplication. However, only the most complete or most recent data should be used in the final analyses, and the duplicates should be highlighted in the flowchart of paper selection.

18.2.2.1 Publication Bias

As one could expect, it has been demonstrated that statistically significant findings have a higher likelihood of being reported than non-significant ones [6–9]. Because of such publication bias, potentially relevant studies could be missing from a meta-analysis.

There are different ways to assess whether publication bias is present in a meta-analysis. The most commonly used methods are based on funnel plot asymmetry [10–12] (Fig. 18.1). In a funnel plot, each study's treatment effect (shown on the x-axis) is plotted against a measure of that study's size or precision, usually using the standard error of the treatment effect on a reverse scale (shown on the y-axis). The name "funnel plot" comes from the fact that the accuracy of the estimate of the effect increases as the sample size increases. Thus, in the absence of publication bias, the studies will be dispersed in a *symmetrically* inverted funnel format. Studies with smaller sample sizes, which lack power and precision, will usually be spread at the bottom. As larger studies are published, the effect estimate tends to remain the same, due to the increasing accuracy, configuring the vertex of the funnel. Nevertheless, there are points of criticism about this method. First, some authors have argued that the visual interpretation of funnel plots is too subjective to be used [13]. Second, other explanations for asymmetry include heterogeneity and methodological anomalies. Finally, as Sterne et al. [14] suggest, the number of studies required to test selection bias by funnel plot should be ten or larger.

Fig. 18.1 Two hypothetical scatter plots of measure of study size vs. measure of treatment effect, known as *funnel plots*. Each dot represents a study. (**a**) Symmetrical funnel plot, suggesting absence of publication bias. (**b**) Asymmetrical funnel plot, with an apparent absence of studies with non-significant hazard ratios (HR ~ 1.0). Adapted from Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001;54 (10):1046–55



18.2.2.2 Gray Literature and Hand-Searching

To minimize the risk of selection bias, it is crucial to find all important data, and also to critically evaluate all existing pieces of evidence, including gray literature, which can be defined as unpublished studies or studies that are not commercially published and, therefore, are not indexed in the relevant databases [15]. In oncology, the more common sources of gray literature are regulatory information (The United States Food and Drug Administration [FDA] and the European Medicines Agency [EMA]), trial registers (clinicaltrials.gov), and conference abstracts. The Scopus and EMBASE databases usually provide unpublished works presented at the main oncology conferences. Other examples of gray literature are book chapters, pharmaceutical company data, letters, dissertations, and theses.

It has been shown that published papers, compared with gray literature, yield significantly larger estimates of the intervention effect [15–18]. Therefore, many argue in favor of including studies from the gray literature in order to more precisely estimate the intervention effect. On the other hand, unpublished studies and studies published in the gray literature lack peer review and might be incomplete, which raises concerns regarding their methodological quality, leading others to question whether they must be included in a meta-analysis. Despite the controversy, the acceptance of gray literature in systematic reviews by researchers and editors is increasing [19, 20] and guidelines for reporting systematic reviews, such as PRISMA [21, 22], AMSTAR [23], and Cochrane [24] recommend that researchers should identify and include all reports, gray and published, that meet the predefined inclusion criteria.

Following the same reasoning as that for searching gray literature, it is suggested that a "hand-search" be performed of the references in the included studies or those in previous reviews. This action can be useful in identifying eligible articles that may not have been retrieved by the search strategy.

18.3 Dealing with Data

18.3.1 Extracting the Data

For most systematic reviews, data collection forms are essential for dealing with published or presented studies. The data collection form is not reported itself, but it is a bridge between what is reported by the original researchers and what is ultimately reported by the reviewers. A good form should include details about the identification of trials, the inclusion/exclusion criteria, risk of bias, methodological aspects of trials, and, finally, data for inclusion in the analysis. Because each systematic review is different, data collection forms will vary across reviews.

It is highly recommended that more than one reviewer extract data from each report, to minimize errors and reduce potential biases that could be introduced by review authors. It has already been shown that, although single data extraction requires less time and fewer human resources, it generates more errors [25]. Special attention should be given to endpoints involving subjective interpretation. Disagreements between reviewers should be recorded and described in the final publication.

When studies are reported in more than one publication or presentation, the data should be extracted from each report separately, and afterward the reviewers should combine information across multiple data collection forms.

Frequently, overall survival (OS) and other time-to-event outcomes (such as progression-free survival or disease-free survival) are evaluated in oncological systematic reviews. These endpoints are best evaluated using the hazard ratio (HR) [26], which is presented with the respective confidence interval (CI). Dichotomous data (such as response rates and adverse events) are usually analyzed using the odds ratio (OR). More rarely the risk ratio (RR) can also be presented.

Sometimes HRs are not presented for OS analyses. However, in almost all cases it is possible to calculate estimates by transcribing the survival curves presented or by using other original data with a spreadsheet developed by Tierney et al. and available online [27]. Continuous outcomes, with mean values and standard variation, are not frequent in oncology trials.

18.3.2 Assessing the Risk of Bias

It is important to understand that, whereas in a clinical study the individual is usually a patient, in a systematic review with/without meta-analysis the individual is a study. Therefore, one pitfall of systematic reviews and meta-analysis is that they are subject to the validity and quality of the studies included. In fact, one can apply a common concept of computer science called “garbage in, garbage out”, where the quality of the output (results from a meta-analysis) is determined by the quality of the input (included studies). Therefore, all studies that meet the eligibility criteria for the systematic review must have their methodological quality assessed on an individual basis. Problems with the design and execution of individual trials raise questions about the internal and external validity of their findings and there is evidence to conclude that biases are introduced into the results of a meta-analysis when the methodological quality of the included studies is inadequate (even when they are randomized controlled trials) [28]. "Study quality" and "risk of bias", will be used here as synonymous, although the Cochrane Collaboration favors "risk of bias" instead of "quality", as "an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research" [24].

The issues of quality assessment are not always related to the design of the trial. Often the trials are just poorly described. In fact, whenever we face an article, we are almost never able to find out how well the study was performed. The only information available for making a judgment regarding a study's risk of bias is the way that it was reported. In other words, we are only able to evaluate how well it was reported. For instance, we are usually not able to evaluate the quality of study procedures, protocol violations, or whether there was any data fabrication or falsification, simply because this information is not usually reported.

Currently, a large number of tools are available for assessing the methodological quality of studies (e.g., the Cochrane tool [29], Jadad [30], and Delphi [31], among others). A meta-analysis may include only high-quality trials; alternatively, a sensitivity analysis (see Sect. 18.4.4) can be done according to the quality of the trials. Each tool has its own instructions, and a detailed description of each one is beyond the scope of this chapter. Items described under the following headings are general concepts of the key methodological subjects often assessed by these tools (discussed more deeply in Chap. 10).

18.3.2.1 Randomization and Allocation Concealment

The included article should report whether randomization was done, and if so, the method used. Random numbers tables, computer random number generators, and stratified or block randomization or minimization are considered to be methods with a low risk of bias. The use of date of birth or date of visit/admission (e.g., even or odd dates) is at high risk of bias. Allocation concealment is responsible for maintaining the effect of randomization in preventing selection bias. The article should report the allocation concealment method. Methods that adequately prevent investigators from predicting the type of group to which the patients were allocated, such as central allocation (e.g., phone, web, or pharmacy), are considered as having a low risk of bias. Trials in which randomization is inadequately concealed are more likely to show a beneficial effect of the intervention [32]. After analyzing 102 meta-analyses that examined 804 trials, of which 272 (34%) had adequate allocation concealment, Wood et al. showed that trials with unclear or inadequate allocation concealment tended to show a more favorable effect of the experimental treatment [33].

18.3.2.2 Blinding/Masking

Low risk of bias means that it is unlikely that the blinding could have been broken or, in the case of an open trial, that the outcome would not be influenced by an inclusion of blinding. Although the lack of blinding has little or no effect on objective outcomes (such as death/OS), it usually yields exaggerated treatment effect estimates for subjective outcomes (such as pain levels). Wood et al. also showed, based on 76 meta-analyses examining 746 trials, of which 432 (58%) were blinded, that intervention effects can be exaggerated by 7% in non-blinded compared with blinded trials [33].

18.3.2.3 Losses to Follow-Up/Exclusions/Missing Data

Incomplete outcome data are due to patient dropouts or exclusions and there are a number of reasons why they occur. It is assumed that the higher the proportion of missing outcomes, or the larger the difference in proportions between the groups, the higher is the risk of bias. Also, there is the theoretical risk that investigators could have excluded patients to favor the experimental intervention. In addition, all randomized patients must be included in the analysis ("intention-to-treat analysis"), which means that a patient who did not receive the intervention, as mandated by protocol, for any reason should not be excluded from the final analysis.

18.3.3 Qualitative Analysis

Although not all systematic reviews have a meta-analysis, they do all have a qualitative analysis, which is presented in the "Results" section of a systematic review. A qualitative analysis usually begins by describing the search process, illustrated by a flow chart, specifying the databases and the number of records retrieved, and giving reasons why studies were excluded. This description gives the reader an idea of the comprehensiveness of the search strategy and increases the internal validity of the review.

It is also during the qualitative analysis that the authors highlight the clinical and methodological characteristics of the included studies, including their size, design, inclusion/exclusion of important subgroups, strengths, and limitations, and the relationships between the study characteristics and the authors' reported findings. All data of interest extracted from each included study, regardless of the number of articles eligible, should be compiled in the form of Tables, making it easier for the reader to have an overview of the studies' main characteristics, including some kind of clinical heterogeneity among the studies.

18.4 Meta-Analysis: Summarizing Results Across Studies

They may seem complex, but all commonly used methods for meta-analyses follow some common principles. Meta-analysis is basically a two-stage process. In the first stage, a summary statistic is calculated for each trial, to describe the observed intervention effect, which is based on the type of variable (Table 18.1). In the second

Table 18.1 Types of variables and their corresponding measures of effect

Type of variable	Effect measures
Dichotomous	Risk ratio (relative risk)
	Odds ratio
	Risk difference
Continuous	Mean difference (difference in means)
	Standardized mean difference
Ordinal	Proportional odds ratios
	Same as dichotomous[a]
	Same as continuous[a]
Time-to-event	Hazard ratio

[a] In practice, longer ordinal scales are often analyzed as continuous data and shorter ordinal scales are often made into dichotomous data by combining adjacent categories together

stage, a pooled intervention effect estimate is calculated as a weighted average of the intervention effects estimated in the individual trials. The weights of each study are chosen to reflect the amount of information that each trial contains, correlated with the sample sizes and dispersion of data; the weights are based on the analysis model (fixed-effect model vs. random-effects model) and the statistical method chosen.

18.4.1 Fixed-Effect Model Vs. Random-Effects Model

The fixed-effect model is based on the mathematical assumption that there is a single common treatment effect (one true effect size) across the studies, and the differences among the effect estimates of each study are attributed merely to chance or type-II errors. If all studies were infinitely large, they would share the same estimates of effect. Therefore, if you consider that all included studies are functionally identical, and have very similar populations and the same experimental and control interventions, a fixed-effect model may be applied. This model will compute the common effect size for this specific population in a more precise manner than the random-effects model, but you should not extrapolate your findings to other populations. This is a rare situation in oncology.

In contrast to the fixed-effect model, the random-effects model assumes that the true effect of the intervention might be different across the studies. This model allows that the true effect size may differ from study to study by chance. This is the reason why the word “effect” is singular in “fixed-effect model” (one true effect) and plural in “random-effects model” (multiple true effects). The random-effects method will usually provide an estimate of the effect with less precision (i.e., with a wider CI), which can be considered a more conservative approach and is indicated in the vast majority of meta-analyses. A recent review of systematic reviews in oncology showed that the random-effects model was underused [34].

Statistically speaking, when using a fixed-effect model, you are pooling together the observed effects from each study (the data you extract from articles) and combining them to make your best guess of what the true common effect they all share really is. Again, if each study was perfect and infinitely large, the observed effects

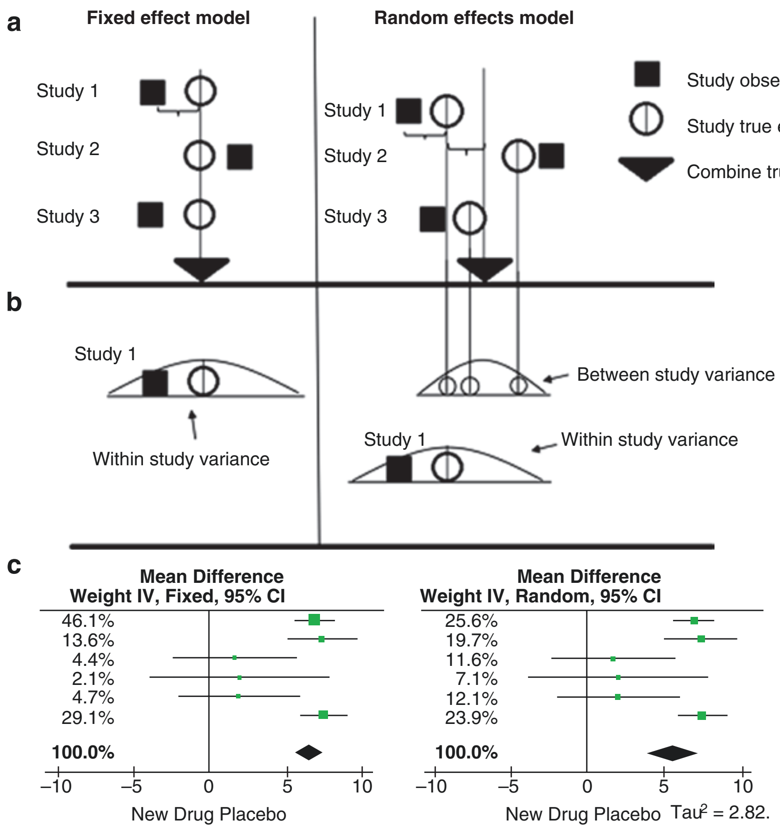


Fig. 18.2 Differences between fixed- and random-effects models. (**a**) The difference between the observed effect (filled square) and the combined true effect has one component in the fixed model and two in the random model. (**b**) This fact leads to one source of variance in the fixed-effect model, while the random-effects model has two sources. (**c**) Example of fixed-effect and random-effects meta-analyses with the same studies. The impact of the method chosen on the weight of each study results in significant differences in the sizes of the squares and the width of the diamonds. Adapted from Borenstein M et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1 (2):97–111

of each study would be the same and equal to the true effect (Fig. 18.2a). The difference between the observed effects in each study from the one common true effect they all share is due only to random errors inherent to each study. Therefore, the fixed-effect model has only one source of variance: the within-study variance. In contrast, in a random-effects model, there are two sources of variance: the within-study variance and the between-study variance. The latter is represented by τ2 (Tau-square). The weight each study receives is (often) the inverse of variance (see Sect. 18.4.2.1). However, while in the fixed-effect model the variance has one component, the random-effects model has two [35]. Therefore, statistically, the only difference between the fixed and random models is how your software weights each study. The weight equals the inverse of the variance in both models, but the variance is further modified by the between-study variance in the random-effects model by using τ2 (Tau-square). Note that, as the meta-analysis shown references in Fig. 18.2c used the random-effects model, the Tau-square was shown (it would be absent in the case of a fixed-effect model).

18.4.2 Statistical Methods

A number of available statistical methods are used to weight effect estimates among the studies included in a review and to pool them together. Three of the most common methods are are outlined below.

18.4.2.1 Generic Inverse-Variance Method

The generic inverse-variance method is one with high applicability because it combines any effect estimates that have the standard error reported. This method can be used to combine dichotomous or continuous data and for fixed- and random-effects models.

Mathematically, variance is the square of the standard error. In turn, standard error describes the extent to which the estimate may be wrong owing to random error. The bigger the sample size of a study, the smaller are both the variance and the standard error. The inverse-variance method assumes that the variance is inversely proportional to the importance of the study; that is, the lower the variance, the more weight will be attributed to this study.

18.4.2.2 Mantel-Haenszel Method

When the data of the studies are scarce in terms of events and/or the studies have small sample sizes, estimates of the standard errors of the effect by inverse variance methods may be poor. In such situations, the Mantel-Haenzel method is preferable, since it uses a different model of weight assignment from that used for the inverse of the variance. This method is used only for dichotomous data, but can be used for both fixed- and random-effects models.

18.4.2.3 Peto Odds Ratio Method

This method is used only for dichotomous data that used the OR as an effect measure and only for the fixed-effect model. It is an alternative to the Mantel-Haenszel method, and is preferable when the two treatment arms have roughly the same number of participants and the treatment effect is small (ORs are close to one) but significant, which is a common situation in oncology.

18.4.3 Assessing Heterogeneity

As the different included studies are not conducted according to the same protocols, they will differ in at least a few aspects. Therefore, a certain level of heterogeneity across studies is usually present, and it can be clinical, methodological, or statistical:

- *Clinical heterogeneity* is due to variability in the included population (e.g., participants' age, performance status, and prior treatments), variability in interventions (different drugs, different dose reduction management of the intervention), and variability in outcome (different definitions of an outcome).
- *Methodological heterogeneity* is due to variability in the risk of bias and/or variability in study design.
- *Statistical heterogeneity* is the variation in the treatment effects of the intervention being evaluated across the studies, i.e., the observed intervention effects are more different from each other than one would expect due to random error (chance) alone. Statistical heterogeneity arises as a consequence of clinical and/or methodological heterogeneity.

Graphically, statistical heterogeneity is presented as CIs from each study with poor overlap. There are statistical tests that can evaluate the heterogeneity between studies. The Chi-square (χ^2, Chi^2, or Q) is one of these tests and it measures how much the difference between effect measures is attributable to chance alone. However, this test has some expressive limitations, such as not being sufficiently powered to detect heterogeneity when few studies are included or when the studies have insufficient sample sizes. Also, as clinical and/or methodological variability often exists [36], some authors argue that detecting statistical heterogeneity could be pointless, since it will be present regardless of whether a statistical test is able or not able to detect it [37]. Therefore, quantifying the heterogeneity may be more useful than simply defining whether it is present or not. The Higgins (or I^2) inconsistency test describes the percentage of variability in the estimate of effect that is attributed to heterogeneity rather than chance. There are different recommendations on how interpret the result of an I^2 test. We suggest the following [37]:

- 0–25%—mild, acceptable heterogeneity
- 25–50%—moderate heterogeneity
- > 50%—high heterogeneity.

When heterogeneity is found, the authors have some options to deal with it:

- Use sensitivity analysis, subgroup analysis, or meta-regression.
- Do not perform a meta-analysis. The authors should only combine studies that are similar enough to be comparable. Although such decisions require qualitative judgments, when heterogeneity is significant and cannot be explained by any sensitivity analysis, the performance of a meta-analysis is not recommended.

18.4.4 Sensitivity and Subgroup Analyses and Meta-Regression

Sensitivity analysis involves repeating the meta-analysis after removing one or a few studies that met the included criteria. Any source of heterogeneity can be the subject of sensitivity analysis to explore its possible influence on the estimation of the effect. Also, sensitivity analysis can be done to find the source of statistical heterogeneity. It is also particularly useful for dealing with outliers, which often overestimate the effect of the intervention.

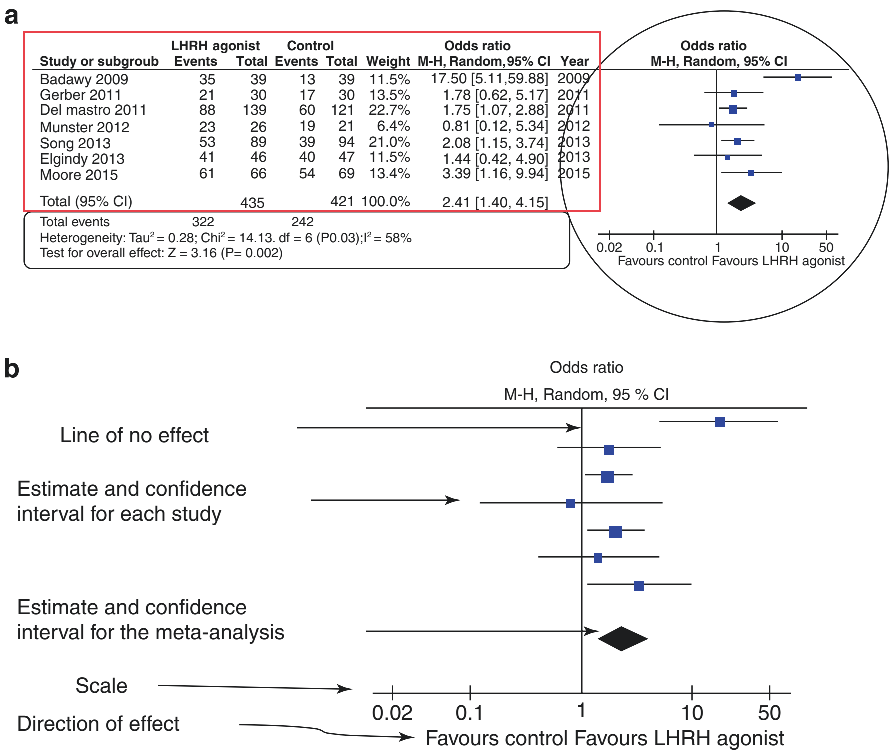
Subgroup analysis involves dividing studies, or the studies' participants, into subgroups according to clinical or methodological characteristics they share. Subgroup analysis of subsets of participants is almost always only possible in individual patient data meta-analysis (see Sect. 18.5). Although each subgroup can be more homogeneous than the entire group, the reader must be aware that subgroup analysis has limitations. First, it decreases the power of the analysis, since each subgroup has fewer studies and patients than the total of the subgroups, which can lead to a false-negative result in a subgroup. Second, the higher the number of subgroups analyzed, the greater will be the likelihood that one of them yields false-positive results. Finally, splitting patients from different studies into subgroups is not based on randomized comparisons, i.e., several other variables may be different and not balanced among patients in a subgroup and, hence, the findings may be misleading.

Meta-regression is a statistical test, similar to multiple regression, which aims to predict the effect estimate according to the characteristics of studies. The advantage of meta-regression over subgroup analysis is that the effect of multiple factors that might have modified the effect estimate can be analyzed simultaneously. However, the number of variables that can be considered to explain effect changes is limited by the number of studies available. Because of this, the Cochrane handbook recommends that "meta-regression should generally not be considered when there are fewer than ten studies in a meta-analysis." [24].

18.4.5 Understanding a Forest Plot

The most usual and informative way to present the results of a meta-analysis is in the form of a graph called a forest plot. This presentation shows the effect estimate and the CI for each study and for the meta-analysis, in addition to allowing rapid inspection of the studies' data and the conclusion of the meta-analysis. Different statistical software can yield forest plots with few differences. Also, the same software is capable of generating forest plots with different information, depending mainly on the type of data and the measure of effect used, as well as what the statistician wants to show. However, all forest plots share the same concepts of presentation.

For didactic purposes, we divided our forest plot [38] into three zones (Fig. 18.3a). In Fig. 18.3a, for zone 1, each line corresponds to a study, which is usually identified in the first column by author name and year of publication or



Study or subgroub	LHRH agonist Events	 Total	Control Events	 Total	Weight	Odds ratio M-H, Random,95% CI	Year
Badawy 2009	35	39	13	39	11.5%	17.50 [5.11,59.88]	2009
Gerber 2011	21	30	17	30	13.5%	1.78 [0.62, 5.17]	2011
Del mastro 2011	88	139	60	121	22.7%	1.75 [1.07, 2.88]	2011
Munster 2012	23	26	19	21	6.4%	0.81 [0.12, 5.34]	2012
Song 2013	53	89	39	94	21.0%	2.08 [1.15, 3.74]	2013
Elgindy 2013	41	46	40	47	11.5%	1.44 [0.42, 4.90]	2013
Moore 2015	61	66	54	69	13.4%	3.39 [1.16, 9.94]	2015
Total (95% CI)		435		421	100.0%	2.41 [1.40, 4.15]	
Total events	322		242				

Heterogeneity: Tau^2 = 0.28; Chi^2 = 14.13. df = 6 (P0.03);I^2 = 58%
Test for overall effect: Z = 3.16 (P= 0.002)

a
Odds ratio
M-H, Random, 95% CI
0.02
0.1
1
10
50
Favours control Favours LHRH agonist
b
Odds ratio
M-H, Random, 95 % CI
Line of no effect
Estimate and confidence interval for each study
Estimate and confidence interval for the meta-analysis
Scale
Direction of effect
0.02
0.1
1
10
50
Favours control Favours LHRH agonist

Fig. 18.3 An example of a forest plot, divided into three zones (**a**) for didactic purposes: the top-left zone (1—red rectangle) provides descriptive data from each study; the right zone (2—circle) presents the graphical nature of the information in zone 1, and the bottom zone (3—black rectangle) shows further statistical components of the forest plot. (**b**) Example of meta-analysis forest plot for interpretation. Courtesy of Munhoz et al. Gonadotropin-releasing hormone agonists for ovarian function preservation in premenopausal women undergoing chemotherapy for early-stage breast cancer: A systematic review and meta-analysis. JAMA Oncol. 2016;2 (1):65–73

the study's acronym. The information in the next columns may vary depending on the type of data. In our example we listed the event rates of each study (number of events in the total of patients in the intervention and control groups). If the meta-analysis had analyzed diagnostic tests, for instance, the forest plot could inform you of the true positives and true negatives instead. The second and third columns in Fig. 18.3a show the weight and the effect estimate of each study (the study's result), along with each study's corresponding 95% CI. The weight is related to the area of the squares in zone 2. The study estimate with it's corresponding 95% CI determines the position where the squares are and the width of the line on both sides. Usually, the bigger the square the smaller the lines. The meta-analysis (the overall effect estimate) is the black diamond that appears below the estimates of the included studies, where its edges correspond to its 95% CI. It is related to the last line of zone 1 (shown in bold).

For the interpretation of the graph in zone 2 (Fig. 18.3b), in addition to the information above, it is important to check the scale, where we will usually find the direction of the effect. Here, studies that concentrate the black squares to the left of the solid vertical line (the line of no effect) indicate results in favor of the intervention and the studies that concentrate their black squares to the right of the vertical line indicate results in favor of the control group. The same applies to the interpretation of the meta-analysis (diamonds). If the diamonds or lines representing the confidence intervals of each individual study are above the vertical line of absence of effect, the interpretation is that there are no statistically significant differences between treatments, or that the meta-analysis is inconclusive. Note that, when dealing with RRs, ORs, or HRs, the absence of effect is represented as 1. When dealing with mean difference, the absence of effect will be represented as 0 (as in Fig. 18.2c).

In zone 3, the first line simply summarizes the total of events. The last line gives you the *p*-value of the meta-analysis. Note that, as the diamonds do not cross the line of no effect, an overall effect *p*-value <0.05 is expected if the CI is defined as 95%. The second line shows data regarding heterogeneity analysis (see Sect. 18.4.3).

18.5 Individual Patient Data (IPD) Reviews

Rather than extracting data from study publications, the original research data may be available directly from the researchers responsible for each study. Individual patient data (IPD) reviews, in which data are provided on each of the participants in each of the trials, are considered the gold standard in terms of availability of data [39]. IPD minimizes the risk of bias and errors resulting from inadequate censoring. IPD can be re-analyzed centrally and eventually also combined in meta-analyses. On the other hand, IPD is usually more costly and time-consuming to obtain than other data. In addition, sometimes the data of all studies that meet the inclusion criteria cannot be available and analysis of only the available data entails a risk of selection bias.

IPD is particularly useful in oncology, where controversial questions and small benefits from interventions are common, and long-term follow-up for time-to-event endpoints (such as OS) is usually required. Situations where publications analyses are based on evaluable patients (not on all patients randomized), or situations where the published information is inadequate or where more complex statistical analysis is required are also well suited for IPD.

18.6 How to Present a Systematic Review with Meta-Analysis

After conducting all systematic steps, before submitting or presenting a review, it is important to return to the original question, and assess how well it was answered by the found evidence. Usually, it is important to evaluate how important the study design flaws are in the interpretation of the meta-analysis. When further research is needed, some specific suggestions can be made about specific design features (better than a simple call for more data).

To assess the applicability of the results the authors should evaluate the inclusion/exclusion criteria. But it is also important to consider how a specific group would differ from the general population.

Presenting a systematic review with meta-analysis is more than just showing the numbers. We suggest a critical assessment, weighing up the beneficial and harmful effects of the interventions evaluated.

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group presents a tool that helps to rate the certainty of the evidence found and the strength of final recommendations [40, 41]. GRADEpro, which can be found on the web (www.gradepro.org), is free and easy to use for summarizing and presenting information.

A systematic review should summarize the evidence in a clear and logical order. The authors can use a variety of Tables and Figures to present information, but we suggest following the PRISMA statement [21, 22] to improve the quality of reports.

Conclusions

As we have seen, through a rigorous methodological process, systematic reviews and meta-analysis help providers to keep up with the enormous amount of research data, judge the quality of studies, and integrate findings. Systematic reviews and meta-analysis yield greater precision of effect estimates, improve external validity (generalizability), providing consistency of results over different study populations, highlight the limitations of previous studies, and contribute to a higher quality of future studies. However, there are many points where authors should be careful in order to not add bias to their analysis and conclusions. Meta-analysis of randomized controlled trials with homogeneity is considered the highest level of evidence [42], but the situation where large randomized trials contradict a prior meta-analysis is still a field of debate [43–45].

References

1. Mulrow CD. Rationale for systematic reviews. BMJ. 1994;309(6954):597–9.
2. Tebala GD. What is the future of biomedical research? Med Hypotheses. 2015;85(4):488–90.
3. Pereira AA, Rego JF, Munhoz RR, Hoff PM, Sasse AD, Riechelmann RP. The impact of complete chemotherapy stop on the overall survival of patients with advanced colorectal cancer in first-line setting: a meta-analysis of randomized trials. Acta Oncol. 2015;54(10):1737–46.
4. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search EMBASE in addition to Medline? J Clin Epidemiol. 2003;56(10):943–55.
5. Sasse AD, Santos L. Searching LILACS database is irrelevant in systematic reviews in oncology. In: Evidence in the era of globalisation. Abstracts of the 16th Cochrane colloquium. Freiburg, Germany. p. 2008.
6. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291(20):2457–65.

7. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med. 2008;5(9):e191.
8. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008;3(8):e3081.
9. Kicinski M. Publication bias in recent meta-analyses. PLoS One. 2013;8(11):e81823.
10. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.
11. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–101.
12. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455–63.
13. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol. 2005;58(9):894–901.
14. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002.
15. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007;2. MR000010
16. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet. 2000;356(9237):1228–31.
17. Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. Int J Technol Assess Health Care. 2000;16(4):1109–19.
18. Burdett S, Stewart LA, Tierney JF. Publication bias and meta-analyses: a practical example. Int J Technol Assess Health Care. 2003;19(1):129–34.
19. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, et al. Should unpublished data be included in meta-analyses? Current convictions and controversies. JAMA. 1993;269(21):2749–53.
20. Tetzlaff J, Moher D, Pham B, Altman D, editors. Survey of views on including grey literature in systematic reviews. 14th Cochrane Colloquium; Dublin, Ireland. 2006.
21. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
22. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006–12.
23. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.
24. Higgins JPT, Green S (editors). Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The cochrane collaboration, 2011. Available from http://handbook.cochrane.org.
25. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59(7):697–703.
26. Keene ON. Alternatives to the hazard ratio in summarizing efficacy in time-to-event studies: an example from influenza trials. Stat Med. 2002;21(23):3687–700.
27. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007;8:16.
28. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609–13.

29. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
30. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1–12.
31. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998;51(12):1235–41.
32. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol. 2007;36(4):847–57.
33. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601–5.
34. Holmes J, Herrmann D, Koller C, Khan S, Umberham B, Worley JA, et al. Heterogeneity of systematic reviews in oncology. Proc (Bayl Univ Med Cent). 2017;30(2):163–6.
35. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1(2):97–111.
36. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy. 2002;7(1):51–61.
37. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.
38. Munhoz RR, Pereira AA, Sasse AD, Hoff PM, Traina TA, Hudis CA, et al. Gonadotropin-releasing hormone agonists for ovarian function preservation in premenopausal women undergoing chemotherapy for early-stage breast cancer: a systematic review and meta-analysis. JAMA Oncol. 2016;2(1):65–73.
39. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof. 2002;25(1):76–97.
40. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6.
41. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4–13.
42. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. JAMA. 2000;284(10):1290–6.
43. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. Lancet. 1995;345(8952):772–6.
44. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA. 1996;276(16):1332–8.
45. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med. 1997;337(8):536–42.

19 Meta-Research in Oncology

Everardo D. Saad and Rachel P. Riechelmann

19.1 Introduction

Most clinical research can suitably be considered primary research, in the sense that investigators design, conduct, and analyze clinical trials which will be the primary source of information that can later inform medical decisions, drug approval and reimbursement, and the research community in general. However, there is a growing field of activity that consists of "research on research", in the sense that data obtained from primary sources are used to generate a second layer of quantitative information that summarizes what is known from primary sources. This activity is sometimes called "meta-research", a relatively new term in the literature. If this term is taken in a broad sense, systematic reviews and meta-analyses are arguably the most established type of meta-research. Since systematic reviews and meta-analyses are covered elsewhere in this volume, in this chapter we will refer to meta-research in a narrower sense, indicating attempts to summarize features of published or ongoing clinical trials, including various methodological aspects of their design, conduct, or analysis. Using this definition of meta-research, we have previously argued that this research modality is a very fruitful avenue for investigators from low-resource countries and settings [1]. It should be noted, however, that scholars in this field define meta-research as a discipline that aims at evaluating and improving research practices and includes five thematic areas (methods, reporting, reproducibility, evaluation, and incentives); of note, such authors explicitly exclude a single meta-analysis on a specific question of interest from the scope of meta-research [2].

E. D. Saad, M.D. (✉)
Dendrix Research, São Paulo, SP, Brazil

IDDI, Louvain-la-Neuve, Belgium
e-mail: everardo@dendrix.com.br

R. P. Riechelmann, M.D., Ph.D.
Department of Clinical Oncology, AC Camargo Cancer Center, São Paulo, SP, Brazil

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_19

Accepting this more scholarly definition, it will be apparent to the reader that most of the discussion and examples in the current chapter relate to the *methods, reporting*, and *incentives* of clinical research. This is a reflection of our own experience and taste, and not of the relative importance of these five areas, all of which account for important aspects of the scientific method. Finally, we ask the reader to forgive our indulgence in self-citation, which is done for the sole reason that by knowing how things were done in our own projects we may openly describe and suggest practical issues that we find relevant in planning, conducting, and reporting meta-research.

19.2 The Idea for a Meta-Research Study

Ideas for meta-research may come from many sources, but we believe they can suitably be grouped as two main types: problem-driven research and data-driven research. Ideally, as in any type of investigation, meta-research projects would aim at finding answers to questions that have been formulated after reflection about important aspects of the scientific method that require improvement, either because they are incompletely understood, not properly evaluated in terms of sound methodology, or have been neglected in the literature. Such a problem-driven approach could either lead to a completely new research avenue or lead to the continuation of an existing line of research. Given the branching nature of scientific inquiry, the latter is particularly common as a motivation for conducting meta-research. However, it should be acknowledged that some projects have a serendipitous character, in the sense that the idea for them comes from looking at data collected for other purposes—similarly to secondary analyses conducted in population databases or retrospective series, when some incidental and interesting findings arise. In this sense, meta-research does not differ from scientific research in general, where serendipity has always played a major role. The reason we emphasize this point is that we have had the experience of developing meta-research projects for which the ideas did not come from reflection, but rather from working on projects that alerted us to the existence of parallel questions that had not been adequately explored in the literature and had only a marginal connection with the original project. As an example, in an initial attempt to investigate aspects related to the definition of progression-free survival in phase III trials on advanced breast cancer [3], a database was formed that allowed—with additional efforts of data collection and analysis—the investigation of issues related to overall survival and post-progression survival [4]. In addition to this somewhat serendipitous work, the database mentioned above also allowed the investigation of issues related to quality of life in the same setting [5], thus also exemplifying a situation for which projects were natural continuations of previous work.

As for meta-research derived directly from a scientific question, the main motive usually entails an overall critical appraisal of the methodology and/or the reporting of studies on a particular topic. Such analytical evaluation often provides a global figure on the quality and on how clinical research has been conducted in that specific setting. For example, we have undertaken a survey of published phase III non-inferiority clinical trials of cancer-directed therapies for advanced/metastatic solid tumors, aiming at understanding the reasons behind the launching of such trials, as well as their characteristics [6]. While non-inferiority trials should be pursued to investigate either more convenient schedules/routes of administration of new medications, less toxic drugs, or cheaper regimens, i.e., those with similar efficacy, with gains for patients and/or society, we found that many such trials have been conducted to approve "me-too" drugs, using large non-inferiority margins to prove non-inferiority.

19.3 Conducting the Search

As in a clinical trial, in meta-research the inclusion and exclusion criteria should be defined a priori, according to the question(s) being investigated. The units of analysis—primary publications rather than patients—should fulfill the eligibility criteria, because any inference obtained from the study sample will apply to a population of primary sources similarly defined. The search for primary sources of information for meta-research projects follows the same general principles as those used in systematic reviews [7, 8]. For example, familiarity with concepts such as medical subject heading (MESH) terms and the "participants, interventions, comparisons and outcomes" (PICO) strategy, and repeated practice with using the PubMed website and the Boolean logic of operators "AND", "OR", and "NOT", are essential skills for conducting meta-research. Access to additional databases is also desirable, and is sometimes a prerequisite for publication in some journals. Since several of these databases require paid subscription, additional useful—and generally open—sources of primary research for oncologists include the Cochrane Database of Systematic Reviews and the electronic abstracts of the most important society meetings, such as the American Society of Clinical Oncology (ASCO), the American Society of Hematology, the San Antonio Breast Cancer Symposium, and others. Using the ASCO website, for example, we have done meta-research on the geographic origin of cancer research, as assessed in abstracts presented at the ASCO Annual Meetings [9].

Despite the necessary rigor required for setting up a search strategy, arguably a difference exists between the search for a systematic review (with or without meta-analysis) and the search for some projects in meta-research. For systematic reviews, publication bias is always of primary concern, because in most cases the investigation aims at describing and quantifying the effects of interventions. In some meta-research projects, since the aim is to assess practice in the published literature, concern with unpublished primary research is, by definition, absent in such projects, and several journals accept manuscripts in which only the published literature has been appraised. Therefore, it is generally acceptable to conduct the search in a single database, which is often PubMed. Moreover, in many cases the scope may be to assess the most influential journals [4, 5, 10, 11] or even only one journal [12]. In our experience, after the scope of the project has been defined with regard to the main research questions, the most practical way to limit the search for a given project is to define the journals of interest (which may be all the published literature related to the question or the journals that are more influential in the scientific community) and to limit the time period to the latest 10–15 years, according to the question. As an interesting example of the choice between limiting or not the number of journals analyzed, we have independently conducted research on the same topic using all the published literature [6] or only selected journals [13]. With regard to the period of interest, if the scope is to assess time trends in the literature, even shorter periods of time may be analyzed, especially if no constraints are applied to the journals of interest [14]. In our experience, limiting the search to English-language primary sources is useful and acceptable, and arguably makes no material difference, as most published clinical trials appear in English-language journals.

19.4 Setting Up a Database

The importance of spending time to set up an adequate and useful database cannot be overemphasized. As usual in research, collecting information that is not insufficient and, at the same time, not excessive is the chief concern. Regardless of the software used (Microsoft Excel being of course the most popular), it is important to organize the database in such a way as to facilitate data retrieval in a reliable manner. Individual sources of primary research must be unambiguously recorded (for example, using the PubMed unique identifier [PMID]) and related sources usually need to be identified when more than one publication has arisen from a given primary clinical trial. Ideally, two investigators should undertake data collection independently, with discrepant cases resolved by consensus or by a third person.

Each meta-research question defines the relevant data to be gathered from articles, and the questions have to be thoroughly defined prior to collection. For example, common variables collected from publications are the number of participants, geographic region, and clinical outcomes. Yet the definition of study variable should not be taken for granted. The number of subjects can be defined as either the number of randomized patients or the number of those who completed the study protocol; the geographic region of a study can be determined by the country of the first author, country of at least half of the authors, or by the country where the majority of patients were accrued. In terms of clinical outcomes, meta-researchers have to be aware of study endpoint definitions, since standard oncological endpoints, such as progression-free survival, may vary across trials [3]. Because more than one researcher generally gathers information from publications, it is crucial that those involved in the study understand the definitions of variables to be collected and be trained before starting data extraction; hence, we suggest that a “dictionary” of variables be elaborated as part of the meta-research protocol to ensure data accuracy.

19.5 Results and Discussion

The results of meta-research are reported analogously to those for any other type of clinical research. The logical sequence is to start with a flow chart of eligible publications, describing the number of excluded articles (and reasons, if possible) and the final number of studies under analysis. This chart resembles those utilized in systematic reviews [7]. Next, there should be a summary of the characteristics of the study "population" of articles (number of patients, type of tumor, oncological setting [i.e., adjuvant vs metastatic], etc), often depicted in a Table. Then investigators describe their findings on the primary and secondary endpoints. For example, in a cross-sectional survey on the proportion of randomized phase II trials that used inferential statistics to report differences between study arms, we observed that either *p* values or confidence intervals had been used by 89% as a form of statistical comparison; as a secondary endpoint in that study, we tested the predictive features of phase II trials that could be associated with the use of any statistical comparison [14].

In meta-research, descriptive statistical analysis is almost always feasible, provided that at least a few primary sources have been retrieved. In addition, meta-research questions are often amenable to inferential statistical analysis, but caution is needed when interpreting these results. The statistical framework of reaching conclusions for populations based on observations made in samples may be present in some cases, but not in others. The very notion of a population parameter may be blurred in some settings, and the inclusion of primary sources that have been retrieved after the use of constraints (with regard to journals and time periods, for example) may preclude unbiased interpretation. In such cases, the use of *p* values, confidence intervals, and statistical modeling should be kept to a minimum.

On the other hand, inferential statistics can be used when the variables and study endpoints are objective and considered to be accurate. For example, we conducted a meta-research of the frequency of self-reported financial conflicts of interest by authors of randomized clinical cancer trials and related editorials and how such conflicts influenced authors' interpretation of study results [11]. To investigate whether there were biased interpretations according to the presence of conflicts of interest, we performed a logistic regression multivariable analysis of factors potentially associated with a more favorable conclusion by study authors. This predictive statistical model could be performed because we transformed a rather subjective variable, "author conclusion", into a categorical variable when we rated authors' conclusions on a scale of "positivity" and grouped them into either positive/highly positive vs. neutral/negative/highly negative conclusions; the scale was pilot-tested for accuracy prior to data collection. Additionally, we selected objective and precise independent variables (type of sponsorship [for-profit vs. not-for profit vs. mixed), primary endpoint result (positive vs. negative) and type of primary endpoint (overall survival vs. surrogate survival outcome variables vs. patient-reported outcomes) to be tested in the multivariable model.

Unlike in clinical trials, in meta-research a formal sample size calculation to estimate the number of articles to be retrieved and included is never performed.

However, if statistical comparisons are planned, e.g., multivariable models, a sufficient number of articles is needed for investigators to properly run the analyses (see Chap. 5).

The Discussion is where investigators critically appraise the literature in the context of their meta-research findings. The Discussion generally follows the regular flow of medical research, where a summary of study results, together with the critical and contextualized evaluation of the literature, is presented. Because meta-research projects tend to evaluate the macro universe of a specific topic, e.g., the overall quality of methods, analyses, and reporting by other studies, they represent a very important tool for scientific advance. In this context, new recommendations and changes in practice can be made. For example, a survey conducted more than a decade ago demonstrated that the quality of reporting of ASCO abstracts of randomized trials was substandard, a finding that caused this society to set guidelines for abstract production and submission [15]. Additionally, meta-research may potentially help clinicians with treatment decisions, as, for example, when transparency is questioned in registration clinical trials that underreport drug-related toxicities [16].

19.6 Analytical Issues and Decisions

Authors of meta-research often need to make decisions about various aspects related to the eligibility of primary sources, the definitions of various concepts, and the best way to analyze the data. It is often necessary to decide which primary source to use for the collection of data for clinical trials for which updates have been published. In some cases, the original publication may be the primary source of information—for example, when the focus of investigation lies in clinical-research methods; in other cases, a subsequent publication (or even more than one) may be more adequate—for example, when long-term results are being analyzed or when the information is present in more than one source. A minor, but recurrent problem exists—when the search is limited by date and is performed in PubMed—for articles published online ahead of print. In these so-called "Epub" cases, a decision needs to be made regarding whether the date of the online publication or the date of the definitive publication should be the one that ensures eligibility for a given study.

When universally accepted definitions exist, they should be used in meta-research. However, it is often the case that ad-hoc definitions need to be created. For example, in some projects it may be of interest to asses phase III trials with regard to certain methodological issues. It sometimes happens that randomized trials are published without an explicit definition of their phase of development. In these cases, we have arbitrarily excluded, for example, clinical trials with fewer than 100 patients randomized or analyzed per arm [5]. Likewise, in some cases the primary endpoint is not explicitly stated by the primary investigators, in which case a definition needs to be created—for example, defining the primary endpoint as the one used for sample-size calculation or first cited in the Results section of the primary article [4].

Table 19.1 Check list for conducting and reporting meta-research

Characteristic	Description
Research question	Should be scientifically sound and relevant, and currently uncertain
Search	Should include descriptions of how the search will/be was performed: database, time span, any language restriction, type of journal, etc.
Eligibility criteria	Explicitly stated and appropriately related to the meta-research question
Endpoints	Clearly stated and defined (e.g., binomial, continuous variable)
Data collection	Should present which data will be/were collected and how the variables will be/were defined
Statistical plan	Should contain details of how data will be/were analyzed according to the type of study variables If statistical modeling is planned, describe the dependent and independent variables
Results	Should include a flow chart of eligible vs. included publications Should include a summary Table of the characteristics of publications Should include reporting of primary and secondary endpoints
Discussion	Contextualization of the meta-research findings to the current literature Evaluation of internal and external validities Discussion of limitations Potentially propose recommendations for change

The principle of equipoise permeates all clinical research, and this is not an exception in meta-research, despite the fact that human beings are not directly involved. In this regard, meta-researchers have to formulate questions based on hypothesis rather than on certainty. This is somewhat obvious, but may lead to unfair conclusions if investigators conduct a biased search and selection of publications in an attempt to prove their own views, e.g., poor-quality studies or studies published from a specific geographic region. Hence, readers have to be critical when interpreting the results of meta-research studies. For this purpose, we recommend a "check-list" of principles that should be included in meta-research projects (Table 19.1).

Conclusions

Meta-research is a useful tool, and one that allows investigators to perform research with relatively scanty resources. Because meta-research studies often evaluate the overall picture of a given topic, they might be published in high-impact journals. Perhaps more than in other fields of research, the idea for the study is more important than the availability of material resources, and ethical constraints do not usually apply in the same manner as for primary research. For all these reasons, we believe that investigators with an interest in clinical research can profit from performing meta-research projects, which, by themselves, are a very efficient (and fun) means to learn more about the methods of clinical trials and principles of biostatistics.

References

1. Saad ED, Katz A, Riechelmann R. Collaboration, niche research and meta-research: three ingredients to innovate and increase the influence of research from Brazil on a global level. Rev Bras Oncologia Clínica. 2011;7((26)):36–46. http://www.sboc.org.br/sboc-site/revista-sboc/pdfs/26/artigo4.pdf (Accessed 08 July 2017)
2. Ioannidis JP, Fanelli D, Dunne DD, Goodman SN. Meta-research: evaluation and improvement of research methods and practices. PLoS Biol. 2015;13:e1002264.
3. Saad ED, Katz A. Progression-free survival and time to progression as primary end points in advanced breast cancer: often used, sometimes loosely defined. Ann Oncol. 2009;20:460–4.
4. Saad ED, Katz A, Buyse M. Overall survival and post-progression survival in advanced breast cancer: a review of recent randomized clinical trials. J Clin Oncol. 2010;28:1958–62.
5. Adamowicz K, Jassem J, Katz A, Saad ED. Assessment of quality of life in advanced breast cancer. An overview of randomized phase III trials. Cancer Treat Rev. 2012;38:554–8.
6. Riechelmann RP, Alex A, Cruz L, et al. Non-inferiority cancer clinical trials: scope and purposes underlying their design. Ann Oncol. 2013;24:1942–7.
7. Egger M, Smith GD, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001.
8. Vincent B, Vincent M, Ferreira CG. Making PubMed searching simple: learning to retrieve medical literature through interactive problem solving. Oncologist. 2006;11:243–51.
9. Saad ED, Mangabeira A, Masson AL, Prisco FE. The geography of clinical cancer research: analysis of abstracts presented at the American Society of Clinical Oncology Annual Meetings. Ann Oncol. 2010;21:627–32.
10. Bariani GM, de Celis Ferrari AC, Precivale M, et al. Sample size calculation in oncology trials: quality of reporting and implications for clinical cancer research. Am J Clin Oncol. 2015;38:570–4.
11. Bariani GM, de Celis Ferrari AC, Hoff PM, et al. Self-reported conflicts of interest of authors, trial sponsorship, and the interpretation of editorials and related phase III trials in oncology. J Clin Oncol. 2013;31:2289–95.
12. Riechelmann RP, Wang L, O'Carroll A, Krzyzanowska MK. Disclosure of conflicts of interest by authors of clinical trials and editorials in oncology. J Clin Oncol. 2007;25:4642–7.
13. Saad ED, Buyse M. Non-inferiority trials in breast and non-small cell lung cancer: choice of non-inferiority margins and other statistical aspects. Acta Oncol. 2012;51:890–6.
14. Saad ED, Sasse EC, Borghesi G, et al. Formal statistical testing and inference in randomized phase II trials in medical oncology. Am J Clin Oncol. 2013;36:143–5.
15. Krzyzanowska MK, Pintilie M, Brezden-Masley C, et al. Quality of abstracts describing randomized trials in the proceedings of American Society of Clinical Oncology meetings: guidelines for improved reporting. J Clin Oncol. 2004;22:1993–9.
16. Seruga B, Templeton AJ, Badillo FE, et al. Under-reporting of harm in clinical trials. Lancet Oncol. 2016;17:e209–19.

Analysis of Health-Related Quality of Life and Patient-Reported Outcomes in Oncology

20

Bellinda L. King-Kallimanis, Roxanne E. Jensen,
Laura C. Pinheiro, and Diane L. Fairclough

20.1 Introduction

Clinical outcomes, such as overall survival, have often been the primary focus of cancer clinical trials and research. However, all treatments have an impact on the quality of patients' lives, symptoms, and functioning, and in 1996, the American Society of Clinical Oncology acknowledged this by noting that patients' quality of life should be a key treatment outcome [1]. Yet the strategies for inclusion of patient-reported outcomes (PROs) to measure concepts such as quality of life are not always consistent (e.g., missing data, poorly defined study protocols) and endpoints or hypotheses tend to be exploratory in nature. Because sample size is most often calculated to power the primary hypothesis, studies are not always powered sufficiently to detect significant statistical and clinical differences for PROs, leading to mixed findings that can be difficult to interpret. With planning and thoughtful execution many of these limitations can be overcome. In this chapter, we highlight key examples of such work.

The additional value of including PRO measures depends on the quality of the instruments used, the quality of the data collected, and the appropriateness of the statistical analyses. In this chapter, we will briefly introduce how to cover these

B. L. King-Kallimanis, M.Sc., Ph.D. (✉)
Pharmerit International, Boston, MA, USA
e-mail: bellindak@gmail.com

R. E. Jensen, Ph.D.
Department of Oncology, Georgetown University, Washington, DC, USA

L. C. Pinheiro, M.P.H., Ph.D.
Division of General Internal Medicine, Weill Cornell Medicine, New York, NY, USA

D. L. Fairclough, Dr.P.H.
Department of Biostatistics and Informatics, Colorado School of Public Health,
Aurora, CO, USA

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_20

important issues pertaining to the quality of patient-centered evidence and provide examples, along with a comprehensive set of references.

20.2 Patient-Reported Outcomes (PROs)

In the literature a number of terms are used for reporting self-reported symptoms, functioning, or quality of life results. As Patrick et al. aptly put it, “the term “PRO” is often used to refer to the things being measured (i.e., concepts and domains (discrete concepts within a multidomain concept)), the instrument used to measure the concepts, and the actual endpoints (i.e., the outcomes as analyzed in a particular clinical trial)” [2]. In this chapter we will generally speak of PRO measures, but when discussing specific concepts or instruments we will use the concept (e.g., health-related quality of life (HRQoL), fatigue, pain, etc) or instrument name (e.g., European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30).

The first step in including PROs in a study is determining the concept of interest, based on the purpose of the study. For example, do patients receiving treatment A have less fatigue than patients receiving treatment B? The next step is reviewing the literature to learn what instruments have been used in past research to measure the concept of fatigue, and in what patient populations. This will yield a list of instruments to consider, and the next step is identifying an appropriate instrument that is both reliable and valid.

Below we outline some broad categories of PRO measures. Examples of common instruments currently being used in oncology studies are provided.

20.2.1 Common Types of HRQoL Measures

These PRO measures capture aspects of a patient’s perceived well-being (i.e., physical, functional, emotional, social, and symptoms). For multi-item concepts, a total score is calculated from the items that have been determined to capture the specific concept. A total score for, say, fatigue, reflects a patient’s relative fatigue as compared with the fatigue of other individuals and should be or compared with the fatigue of the same patient at follow-up assessment points. These PRO measures can be broken down into two subtypes: generic and disease-specific PRO measures.

20.2.1.1 Generic Instruments

Generic instruments allow researchers to compare across disease groups and cancer types, and to compare with the general population, as the items are not specific to any one health condition. For example, the SF-36 [3] is widely used to assess HRQoL and includes 36 items and 8 subscales: Physical Functioning, Role-Physical, Role-Emotional, Bodily Pain, General Health, Vitality, Mental Health, and Social Functioning. The SF-36 also includes Physical Component Summary and Mental Component Summary scores, which are weighted summations of all items. There are also shorter versions (i.e., short forms) of the SF-36, such as the SF-20 and SF-12, which are attractive to use when patient time constraints are of particular concern.

Other types of generic instruments are preference-based measures. These instruments are influenced strongly by the concept of utility, borrowed from econometrics, and they reflect individual decision-making under uncertainty. Preference-based measures derive a single value to represent the HRQoL associated with a given health state. In general, a value of 0 represents death and a value of 1 represents perfect health. The most widely used preference measure is the EQ-5D questionnaire developed by the EuroQOL group [4]. On the EQ-5D, patients are asked to rate "How good or bad is your health today" on a visual analog scale (VAS), a 100-mm continuous line which ranges from 0 to 100. In addition to the VAS, five dimensions of functioning are assessed, using either a 3- or a 5-point response scale. These profiles create health states that can be used to calculate a utility score. These utility scores come from using a value set that is available on the EuroQol website [5], where calibrated value sets are available for 20 countries. Utility scores can also be used to estimate quality-adjusted life years (QALYs) for assessing the cost-effectiveness of different drugs or treatment modalities.

20.2.1.2 Disease-Specific Instruments

A disease-specific instrument allows researchers to capture concepts that are specific to a defined patient group. For cancer patients, this can allow one to assess across all cancer types or within a site-specific type. For example, a general cancer-specific PRO item may ask, "Were you short of breath?" [6], whereas a lung cancer-specific PRO item may ask, in addition, "Did you cough up blood? [7]". The choice to include either general cancer items or cancer-site specific items will be determined by how sensitive the researcher believes the measure will be to capture differences or changes in the symptoms assessed and whether the more specific symptoms are relevant to their sample.

Two frequently used "gold standard" PRO measures in cancer are the EORTC QLQ-C30 and the Functional Assessment of Cancer Therapy—General (FACT-G). The EORTC QLQ-C30 covers nine multi-item domains, five of which are functional scales (Physical, Role, Cognitive, Emotional, and Social); three multi-item symptom domains (Fatigue, Pain, and Nausea and Vomiting); six single items (Dyspnea [shortness of breath], Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties); and a Global Health and Quality of Life domain [6]. The FACT-G includes four domains: Physical Well-Being, Functional Well-Being, Emotional Well-Being, and Social/Family Well-Being [8]. Both instruments have been widely used in cancer clinical trials and research studies.

A full discussion of the strengths and weaknesses of ways of choosing one of these instruments is beyond the scope of this chapter; however, Luckett et al. (2010) wrote a review including a decision tool to help in deciding whether to use the QLQ-C30 or FACT-G [9].

Cancer site-specific modules have also been developed for the EORTC QLQ-C30 and FACT-G instruments. The general core items are presented for all patients, along with the site-specific module. These modules are briefly described below.

20.2.1.3 EORTC QLQ-C30 Modules

The cancer site-specific add-on modules focus on symptoms and functional issues commonly identified by patients diagnosed with a particular cancer type. In general, a total score is created, using the module items, by following the published scoring algorithm. The modules cover a range of cancer types, disease severities, and treatment modalities. The Head & Neck module (QLQ-H&N35), which consists of items such as "Have you had pain in your jaw? [10]" and the Breast module (QLQ-BR23), which asks items such as "Was the area of your affected breast oversensitive? [11]" are frequently used in research. Currently, over 22 modules in various stages of development are available for use at the EORTC website [12].

20.2.1.4 FACT Modules

The site-specific domains in these modules include concepts specific to a particular cancer type, specific symptom, or treatment. For example, the FACT-B (Breast Cancer), includes a 10-item "other concerns" subscale covering content specific to breast cancer (e.g., body image, hair loss, and feeling sexually attractive) [13]. In contrast, the FACT-L (Lung Cancer) "other concerns" subscale covers topics such as multiple coughing items, and regret about smoking. Site-specific FACT scores are summed as a separate subscale and added to the FACT-G.

To date, 48 site-specific validated modules and short forms (i.e., instrument versions with fewer items) are available for use. These include instruments for both adult and pediatric populations, and have equivalent non-English-language versions. Additional information regarding these PRO measures can be found on the FACIT website, which includes a number of PRO measures beyond cancer [14].

20.2.1.5 Measure Selection and Scope

Disease-specific PRO measures allow researchers to measure concepts that are most relevant for a patient population. For example, cancer patients are more likely to identify symptoms (e.g., feeling weak) that are not present for, say, patients with overactive bladder. This is important in a drug trial to demonstrate superiority in the treatment arm and for health technology assessment submissions where reimbursement and coverage decisions may be made. Disease-specific items also allow for a closer focus on the patient population of interest, and a reduction in ceiling and floor effects, where a large number of patients (100/number of response options on an item) pick the highest or lowest response option across items. However, a disadvantage of disease-specific measures is their lack of generalizability or ability to interpret what scores mean beyond the population being measured (i.e., other cancers, other diseases, the general, non-cancer, population).

Limitations: Using cancer-specific PRO measures does not allow for comparisons against the general population without cancer or comparisons with other patient populations (e.g., those with other chronic diseases). Cancer-specific PRO measures may not have been validated across as wide a spectrum or geographically diverse patient populations as general measures. In interpreting disease-specific scores, there may be no published work describing what a minimal important difference or meaningful change equates to.

Ideally, the inclusion of both a general and a cancer-specific PRO instrument would provide complementary information. By including a general PRO measure, a comparison of PRO scores can be made with the scores of the general population. This is important when trying to determine the clinical meaningfulness of a particular instrument's scores. Some disease-specific instruments such as the FACT-G have healthy general population normed scores, which can be used as a reference when interpreting FACT-G scores in a specific cancer population [15]. Finally, when exploring the added value of new therapies, general measures enable comparisons across indications.

Of the PRO measures discussed in this section, the choice of which one to include in a study depends on the research question. In clinical trials, the choice also depends on the payers and regulators and their requirements for an application of a new drug or device. As PRO measures have been developed to answer targeted research questions, alternative PRO measures have also been developed. Below is a brief description of some of these alternatives.

20.2.1.6 Single Items and Computerized Adaptive Tests (CATs)

The patient burden of completing additional multi-item PRO measures is an important consideration when selecting the appropriate instrument to use in a study. There is debate in the literature around the ideal number of items required to measure a concept such as HRQoL [16]. In some very specific cases when resources are extremely limited or with advanced-stage patients who tend to be frail, a single item may suffice for measuring a concept of interest. There will be a loss of richness to the data (e.g., no subgroup analyses based on domains) as most concepts are complex and require multiple items. However, depending on the research question, conclusions drawn from a single item may not drastically change results [17].

Computerized adaptive testing (CAT) is an administration method for Item response theory (IRT)-calibrated item banks and pools, that minimizes the number of items, while maximizing the measurement information from each item [18]. CAT selects items based on a patient's response to the prior item. For example, in an assessment of physical function, if a patient answers that he/she can "never" walk a mile, the next physical function item will represent a lower "difficulty" of physical function, such as the ability to walk a block. This person will never be asked if they can run a marathon, as there is a high probability this person would endorse "never" [19]. CAT administration requires an electronic administration.

20.2.1.7 Symptom Severity, Adverse Events, and Performance Status PRO Measures

Clinician-reported measures primarily capture disease symptom severity, treatment-related adverse events (AEs), and performance status (e.g., Eastern Cooperative Oncology Group [ECOG]). However, the reliability of clinician-reported measures has consistently shown poor agreement with patient self-reports [20]. Two recent efforts to systematically capture a broad, comprehensive range of patient-reported AEs and symptom severity are briefly described below.

PRO-CTCAE: The patient-reported version of the Common Terminology Criteria for Adverse Events (CTCAE) is an item library including 124 PRO items that capture 78 distinct symptomatic AEs reported during cancer treatments. All PRO-CTCAE items use a 7-day recall period and a 5-point Likert-type scale, except where absence or presence is assessed. Symptomatic AEs are assessed for one or more of the following attributes: (1) Absence/Presence, (2) Frequency, (3) Severity, and (4) Interference with Daily Activities.

PROMIS®: The Patient-Reported Outcomes Measurement Information System (PROMIS) is a set of PRO measures that cover a range of symptoms and functional domains. For each domain, an IRT-calibrated item bank can be used to create short forms or CAT item administration. Regardless of the administration method, instrument length, or patient population, PROMIS scores are comparable on a single *T*-score metric anchored against the United States general population, and have United States reference values for many cancer populations [21].

The PROMIS Profiles 29, 48, and 52 include short forms (4, 6, and 8 items, respectively, per domain), with the domains being Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Satisfaction with Social Roles, and Pain Interference, and one single item on Pain Intensity.

Both the PROMIS and PRO-CTCAE measures are currently being used, particularly in some major United States cancer clinical trial collaborative groups (e.g., the ECOG and the American College of Radiology Imaging Network [ECOG-ACRIN]).

20.2.1.8 Individualized Measures

Patient-generated index (PGI): The PGI allows patients to determine the HRQoL domains that are of the greatest importance to them [22]. In clinical trials or large-scale studies, the PGI presents some implementation challenges. For example, responder bias associated with a patient's level of education, and fatigue owing to the more complicated nature of the instrument are of concern [23]. In addition, methods to assess changes in PGI domains over time are not well established for clinical trials. However, the PGI has value in the clinical setting where patients and clinicians can identify HRQoL areas that are of key concern to be further explored [24].

20.3 Measure Selection

20.3.1 Purpose

Research studies, the most common use, should select PRO measures that support a clear objective it is too repetitive to keep repeating PRO measures. The number, timing, and location of assessments are fixed. Study personnel are involved (in person or by phone) in measure administration and follow-up in an effort to reduce or eliminate participant non-response. PRO measures that are used in research studies can be longer and more detailed than those used in clinical trials, to ensure the study is sufficiently powered to report significant outcomes. A check list including best practices for the inclusion of PROs has been created [25] as a Consolidated Standards of Reporting Trials (CONSORT) extension [26]. For example, the CONSORT check list for PRO recommends reporting on whether the PRO is a primary or secondary endpoint; instrument validity; hypotheses of PRO; statistical handling of missing data; and discussion about PRO limitations and external validity for other settings [26]. The use of such a check list can help ensure that appropriate PRO measures are considered, applied, and effectively reported in clinical trials.

In clinical care settings, the administration of PRO measures must be directly relevant to the patients, quick to administer, integrated into everyday workflow, and include clear resources to help providers interpret and use scores. The International Society for Quality of Life Research (ISOQOL) has created a user guide for implementing PROs in clinical practice that addresses these requirements [27].

Once content and purpose have been established, the type of measure should be selected. The first decision to be made is whether there is a “gold standard” for the population of interest or whether a newer measure should be selected or possibly developed (time permitting). For example, both the FACT and EORTC core measures have been long established for use in cancer patient populations. Administration of these measures is straightforward; the measure is administered as a fixed form, and scored by summing items within each subscale, creating an overall total score.

Developed 10 years ago, PROMIS measures are newer, with different administration considerations. PROMIS has both fixed “short forms”, while also allowing for the creation of a customized short form from domain-specific item banks. Scoring requires specific psychometric software (IRTPRO), the use of published tables (established short-forms only), or the submission of scores to a dedicated online scoring service [28]. PROMIS measures have been validated in cancer populations, but are generic PRO measures.

Important Considerations for Measure Selection:

1. Has the measure been shown to be valid and reliable in cancer patients?
2. If so, how broadly has it been tested and used? Are there studies evaluating differences by race/ethnicity, age, and other factors that may change how a patient may answer a question?
3. Is the measure available in other languages? Other formats (electronic, paper)?
4. Is the measure in the public domain? If not, how much will it cost to administer?
5. How long does it take to fill out the form?

20.3.2 Administration Options

Mode of administration (paper, in-person, phone, computer) is always a consideration in study design. However, research suggests that electronic and paper administrations result in no meaningful differences. Gwaltney et al. (2008) performed a meta-analysis of 46 studies and 275 PRO measures to examine relationships between paper-administered PROs and computer screen-based ePROs [29]. The average mean difference between the two modes of administration was small (0.2% of the scale range or 0.02 points on a 10-point scale). The average correlation between the paper and ePRO modes indicated redundancy (0.90). That is, the two modes of administration were highly correlated with each other.

However, electronic administration also requires hardware (e.g., computer, touchscreen tablet), study software, and stable Wi-Fi access. Wi-Fi access has been shown to be challenging in some hospitals, which may lack coverage underground or in shielded areas (i.e., radiation oncology) [30].

20.4 Design

20.4.1 Sample Size

Sample-size estimates for PRO measure analyses are the same as those used for other sample-size estimations. The researcher must identify a hypothesis and determine an appropriate clinically meaningful difference for the study. If the difference is a change from baseline, determinations of what is clinically meaningful may be guided by previously published minimally important differences (MIDs). The amount of expected attrition should also be considered when performing sample-size calculations. For example, if 200 patients will be recruited for a study and typical attrition is 20%, the power calculation should be based on 160 patients.

A number of guidelines have been published to aid in the determination of sample-size calculations by providing estimates of MIDs that can be used for both sample-size calculations and the interpretation of results [31, 32], though these guidelines are not often put into practice. Cocks et al. (2011) published more comprehensive guidelines for the EORTC QLQ-C30. Included in these new guidelines are subscale estimates to calculate sample size based on small, medium, and large effects [33].

20.4.2 Timing of Assessments/Patient Burden

Two important and related considerations for PRO measures are (1) the timing of assessments and (2) patient burden. In open-label or un-blinded intervention studies, assessment of concepts prior to patients learning which arm they have been assigned to is critical to reduce bias associated with being in the treatment arm versus the standard-of-care arm. In longitudinal studies where change over time in a PRO measure is of key interest, careful planning of assessment points is required. Important considerations include the timing of treatment, duration of treatment side effects, disease symptoms, and disease severity. In addition, convenience of assessments is important for patients. For example, assessments, particularly in clinical trials, are often linked to clinic visits, although these may not be optimal times for capturing the concept of interest. Another important consideration is how to capture PRO measures, especially HRQoL, at clinic visits. For example, if the PRO measure is collected when the patient receives information on their disease progression, this may bias responses. That is, patients who complete a PRO measure prior to receiving disease progression information may respond differently from patients who complete the PRO measure after receiving disease progression information. Another issue to consider is learning. Patients who are assessed too frequently for their performance on cognitive neuropsychological tests tend to *learn* the tests if such assessments are repeated at short intervals.

With respect to patient burden, if a patient were very sick (e.g., stage IV pancreatic cancer), it would not be appropriate to burden them with a lengthy questionnaire, owing to risk of drop out. This is of particular concern when studying end-of-life cancer patients. Finally, greater patient burden may also contribute to a greater level of missing data, which is also of concern.

In general, careful consideration of assessment timing and patient burden has the potential to reduce missing data.

20.4.3 Open-Label Clinical Trials

In the case where blinding is not possible, most often owing to differing treatment schedules, there is the potential for bias. Subjects who are aware of their study arm assignment may differentially report their HRQoL compared with individuals who are unaware of their assignment. Therefore, collecting HRQoL data at screening and prior to assignment is important so that shifts can be investigated, as learning the treatment assignment may lead to changes in how the patient views their HRQoL. The impact of open-label trials on PRO measures is an area that has not yet been well researched.

20.4.4 Missing Data

When missing data is present, a study's power is reduced and there is a risk of selection bias, which can impact the internal validity of study findings. In the past, the last observation carried forward (LOF) was the standard practice for handling missing data. While methods for handling missing data have advanced beyond LOF, preventing missing data through study design is still the best solution. Prevention could be ascertained via adequate resources to support the inclusion PROs. For example, studies can ensure that sites receive training on the standardized administration of PRO measures and that the study protocol clearly outlines the implementation strategy and actions for protocol deviances. In addition, selection of research/clinical sites with a track record for excellent compliance can help to reduce missing data. Finally, the selection of a PRO measure and administration mode that is most appropriate for the population and for the setting in which assessments will take place can also reduce the level of missing data. For example, the Swiss Group for Clinical Cancer Research conducted interviews with patients, relatives, and nurses to understand low levels of compliance. All groups had overall positive feelings toward the inclusion of PRO measures; however, there were concerns from some patients that responses could potentially lead to different treatment options and nurses were concerned about research demands versus patient care obligations [34]. The first concern could be remedied by providing a thoughtful explanation to patients of how the data will be used. The second concern could also be addressed through careful considerations of the timing and mode of administration of PRO measures, with the inclusion of all involved team members in this decision-making process. For a comprehensive review of preventative strategies for missing PRO data, see Mercieca-Bebber et al. (2016) [35].

20.4.5 Multiplicity

With the inclusion of PRO measures, the number of statistical tests adds up quickly. This can be due to numerous primary and secondary endpoints, as well as the multiplicity generated by the longitudinal assessments (multiple time points) on multiple scales (e.g., physical well-being, emotional well-being, or multiple symptoms). This leads to the possibility that false-positive results may occur when the number of tests conducted is not considered and accounted for in the evaluation.

If there is only one time point, it is possible to use the Bonferroni approach, a *post-hoc* adjustment method, which involves dividing the level of significance (i.e., type-1 error rate) by the number of tests being conducted. For example, if there are eight domains to be tested at the 5% error rate, then 5% is divided by 8 (i.e., 5%/8 = 0.62%) and the adjusted rate is 0.62%. As this is a conservative adjustment, other Bonferroni-like adjustments (e.g., Benjamini-Hochberg procedure) have also been proposed [36–38].

However, *post-hoc* adjustment is often not feasible because it is impossible to determine the number of tests performed over the course of the study. This approach also makes the strongest assumptions about missing data that may be difficult to justify. Other strategies may include: limiting hypothesis tests to a reduced number of measures pre-specified in the trial or study design, considering summary measures or statistics across time or across subscales, utilizing multiple comparisons adjustments or gate-keeping strategies (i.e., a hierarchical order of objectives is determined and secondary hypothesis testing will be performed only if the primary hypothesis, the gatekeeper, is rejected).

20.5 Types of Studies

20.5.1 Observational Studies

PRO measures can be collected in large observational studies to better understand patterns and trends when randomization is not feasible. In observational studies, researchers are able to gain perspectives regarding concepts in the "real world" or in population-based settings; such perspectives are important when considering large-scale health policy impacts on the concepts important to patients, as well as when considering strategies to scale-up successful interventions from randomized control trials. There are different types of observational studies, including cohort, case-control, cross-sectional, case-crossover, and longitudinal studies, the choice of which will depend on the research question and the resources (i.e., funding and time) available.

As there is no randomization in observational studies, confounding or biased estimates are important to understand and consider when interpreting results. These biases can be handled via the design of the study and the inclusion of specific variables that can be used to stratify groups of patients, or used as covariates in an adjusted multivariable model. Another method to reduce confounding in observational studies is by propensity score matching [39].

In the United States, several large observational studies have included PRO measures that are cancer focused, or that can be used for both cancer and other diseases. Some of these observational studies include: the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) cancer registry linked with data from the Center for Medicare and Medicaid Service's (CMS) Medicare Health Outcomes Survey (MHOS), the Health and Retirement Survey (HRS), and the Medical Expenditure Panel Survey (MEPS). These observational studies collect generic PRO measures such as the SF-36/SF-12, Global Health, Depression, Anxiety and Activities of Daily Living items, and vary in design. With the HRS and MEPS using longitudinal panels, and the SEER-MHOS using both cancer patients and non-cancer controls over time. As such, researchers are able to use these observational studies to examine research questions at one point in time (cross-sectional) and over time (longitudinal).

Using these large observational studies, numerous studies have been conducted to examine PRO patterns and investigate predictors. For example, in one of the first SEER-MHOS studies published, Reeve et al. used SEER-MHOS to document the negative impacts of cancer diagnosis/treatment on HRQoL across the eight domains of the SF-36 in nine cancer types (prostate, breast, colorectal, lung, bladder, endometrial, or kidney cancers; melanoma; or non-Hodgkin lymphoma) [40]. Using this large, population-based, dataset allows for increased generalizability of study findings to the general United States population. In a study using the HRS dataset, investigators found that the mental health of older adults who had survived any cancer diagnosis for 4 years or longer was similar to the mental health of those without a cancer diagnosis. However, physical health (e.g., Activities of Daily Living, Mobility, and Pain), for some survivors, was worse than that of people without a cancer diagnosis [41]. Finally, an observational study using MEPS found that HRQoL among cancer survivors depended on both the time since cancer diagnosis and the cancer type. For example, after 10 years, breast cancer, colorectal cancer, and melanoma patients had HRQoLs similar to those of individuals without cancer, whereas cervical and prostate cancer patients experienced lower levels of HRQoL [42].

20.5.2 Clinical Trials

In clinical trials, PRO data is included to support secondary and exploratory endpoints, as the primary endpoint is focused on assessing the treatment benefit and superiority effect of a new drug on survival or perhaps, progression-free survival. Therefore, the inclusion of PRO measures is not always well integrated in the study design and many of the analyses focus on differences in change from baseline between the treatment arms. However, in advanced disease or assessments of second-line treatment, there may be very little shift in survival for a new drug. In this case, HRQoL improvements and reductions in treatment-related AEs and toxicities may be important considerations for assessing treatment benefit, and therefore the timing of assessments of HRQoL and/or symptoms is critical.

In consideration of symptom assessments, numerous studies have found improved reporting of symptoms and AEs when patients self-report their own symptoms and AEs. Of particular importance are subjective symptoms such as pain, fatigue, and nausea. Subjective symptoms are more difficult for clinicians to be aware of and previous studies have found that clinicians underreport the frequency and severity of subjective AEs in oncology [43–45]. For example, in one study, the agreement between patient- and clinician-rated fatigue was considered low (kappa = 0.07) [46], where perfect agreement would equate to a kappa of 1, and chance agreement would equate to essentially 0.

Before a PRO measure can be included in a clinical trial, it must undergo a thorough evaluation of the instrument’s measurement properties. On the other hand, the clinician-reported CTCAE has not been formally evaluated and it is undetermined as to whether there are important differences for patients between a grade 2 and a grade 3 of some subjective AEs, such as diarrhea and fatigue.

A systematic review of the prognostic impact of PROs on overall survival (OS) found that, in 36 of 39 clinical trials examined, PROs were significantly associated with survival at the 0.05 level [47]. This may be especially relevant when clinical trials fail to identify differences in treatment benefits between study arms. For example, in a secondary data analysis from a study examining treatments for patients with hepatocellular carcinoma (HCC), the authors found no differences in treatment benefits across study arms. In that study, the prognostic impact of HRQoL on OS was also assessed [48]. HRQoL was self-reported using the EORTC QLQ-C-30, and the authors found that the Role Functioning, Fatigue, and Diarrhea domains were independent predictors of OS in these palliative HCC patients.

20.6 Analysis

20.6.1 Methods of Dealing with Missing Data

Missing data is ubiquitous, and a number of methods have been proposed for handling it. However, before a method is chosen, an understanding of the mechanisms underlying the missing data is required. There are three types of missing data: (1) missing completely at random (MCAR), (2) missing at random (MAR), and missing not at random (MNAR) (Table 20.1). MCAR is where the data does not depend on the observed or unobserved outcomes of interest. For example, if a patient goes on vacation and misses an assessment, their going on vacation does not depend on the outcomes of the study. MAR is when the missingness depends on the observed outcome. For example, a patient has progressive disease (an observed outcome) and changes treatments and has missing observations during this time. Finally, MNAR is when missingness is dependent on unobserved outcomes. For example, patients who are highly fatigued do not complete the fatigue PRO measure, and as a result only patients with less fatigue are included in the analysis. In clinical studies, data are infrequently MCAR, and determining whether the data are MAR or MNAR is difficult, as the information required to make the distinction is missing from the study.

When missing data rates and causes of missing data differ between treatment arms, PRO endpoints will be impacted in unknown ways. For clinical trial research, analyses must be explicitly laid out in a statistical analysis plan before database lock where the data is transferred for statistical analysis. As MCAR is an unlikely reason for missing PRO data, methods relying on the MCAR assumption are usually not appropriate.

Analyses for data assumed to be MAR include mixed-model for repeat measurement (MMRM) and mixed-model growth curves (MMGCs) where all observed data are used. These types of analysis assume that, conditional on past history, patients in a specific treatment arm would have had results similar to those of other patients in that treatment arm had they not dropped out of the study or trial. This may or may not be a reasonable assumption, depending on the study. In the European Medicines Agency (EMA) 2010 guidelines on missing data it is stated that "these methods will, in certain circumstances, overestimate the size of the treatment effect likely to be seen in practice, and hence introduce bias in favour of the experimental treatment." [49].

Table 20.1 Definitions of missing data

Assumption name	Acronym	Dependent on	Independent of
Missing completely at random	MCAR	Covariates	Observed outcome Missing outcome
Missing at random	MAR	Covariates Observed outcome	Missing outcome
Missing not at random	MNAR	Missing outcome	–

As it is not possible to assess with certainty whether data are MAR or MNAR, the understanding of missing data patterns is crucial and requires a series of analyses. One method for assessing MNAR is pattern mixture modeling [50, 51]. In this type of model, patients are divided into attrition groups. For example, patients would be grouped in either an early attrition (dropout prior to disease progression), late attrition (dropout post progression), or a completer group (completed all assessments). Within each attrition group, least square mean estimates of average change in each study outcome from baseline, along with 95% confidence intervals, would be calculated. If attrition patterns vary significantly by treatment arms, PRO analyses are often stratified by attrition group in order to reduce bias. Understanding the missing data patterns is key to choosing the best way to handle the missing data.

20.6.1.1 Missing Data and Death

In oncology studies and trials, missing data due to death is an important issue for consideration. The preference-based measures define death as equal to zero, but other PRO measures rarely assign a value for death. Imputation can be avoided by using primarily descriptive statistics to graphically depict the domains of interest by strata, but the interpretation is limited to just that and cannot be used as an overall comparison of the outcome under the intent-to-treat concept.

Imputations have been suggested, for example, imputing a high or low value that falls outside of the range that would be expected by living patients [52, 53], or a zero could be imputed, as is done for the EQ-5D. None of these strategies is entirely defendable, and if a substantial proportion of patients die, the distribution of scores will become bimodal. This is because the score would be approximating a binary indicator for death, and the analysis would become an approximation of the analysis of survival rather than the outcome of interest.

Pattern mixture modeling can also be used for modeling missing data due to death. This would involve modeling a stratum defined by time of death. A review of unconditional and conditional models for longitudinal data truncated by death investigated both unconditional (data were implicitly imputed) and conditional models. The conclusions drawn regarding the domain for cognitive functioning were found to be dependent on the method used to estimate the model [54].

Another approach is to jointly model the longitudinal changes in the PRO measure with time to death (or progression). The underlying assumption is that the data are MAR, conditional on the time to the event. The notion is that the random effects of the model for the PRO outcomes are correlated with the time to the event. Specifically, patients who experience an earlier event will tend to start with lower scores and will decline more rapidly. A wide variety of parametric and non-parametric models [55–57] has been proposed. Vonesh et al. (2006) [58] proposed extending the model by

1. Relaxing the assumptions of normality.
2. Allowing distributions of the random effects from the quadratic exponential family.
3. Event-time models from accelerated failure-time models (e.g., Weibull, exponential extreme values, and piece-wise exponential models).

Numerous investigators have joined proportional hazard models with the longitudinal models. Other extensions include multiple reasons for dropout [59, 60] and the possibility that some subjects would not eventually experience the dropout event and could stay on the intervention indefinitely [59, 61, 62].

In oncology studies, missing data due to death will almost always be an important consideration; matching the aims of the study to the appropriate analysis methods is key to taking this consideration into account.

20.6.2 Longitudinal Data

20.6.2.1 Mixed-Effect Models

Two types of mixed-effect models are well suited to longitudinal analyses because an adjustment is made for the correlation between repeated observations within the subject. The choice between a growth curve mixed model (GCMM) [63] and a repeated-measures mixed model (RMMM) [64] will depend on the timing and number of assessments. Trials with a limited number of assessments (two to five) that can be thought of as ordered categories (e.g., Pre, Early, Late, and Post Therapy) and where all assessments can be uniquely classified are typically analyzed using a repeated-measures model. Trials with a larger number of assessments or those where the timing of assessments becomes more varied over time are typically analyzed using GCMMs that incorporate both fixed effects (i.e., regression coefficients) and random effects (i.e., variance around the regression coefficients). A simple mixed-effect model in a clinical trial assessing a PRO measure will include time, treatment arm, baseline PRO measure (fixed effects), and a treatment arm-by-time interaction (random effects). When using these models, the underlying assumption is that any missing data are MAR. If, for example, a decline in HRQoL occurs after dropout, then results from these models will be biased, as discussed in relation to missing data.

Mixed-effect models are favored over, say, an analysis of variance (ANOVA), where imputation of missing data is required when assessing mean change from baseline to an endpoint. In clinical trial analyses, most often, the data were imputed in the past using LOF. A simulation study that used 25 new drug application datasets from clinical trials found that mixed-effect models were superior to LOF with ANOVA in controlling type-I error rates and minimizing biases [65].

20.6.2.2 Time Until Deterioration

Time until HRQoL deterioration (TUD) is a statistical technique for analyzing longitudinal HRQoL data [66, 67] where time-to-event analyses are used. The threshold for a clinically significant deterioration in the HRQoL domain of interest is defined a priori. For example, a deterioration of 10 points on the domains of the EORTC QLQ-C30 has been used in the past, as Osoba et al. (1998) proposed that a 10-point deterioration for this instrument mapped to a minimum important difference (MID) for patients [68]. The threshold used to define deterioration should be clinically meaningful in relation to the study and patient population under consideration. Death can be included as a competing event to deterioration, as can missing data. Patients who did not experience a drop in HRQoL by the last follow-up are censored.

Furthermore, how time is defined is an important consideration. In the adjuvant setting, time from baseline (e.g., randomization) to the first clinically significant decrease in HRQoL may be the most appropriate definition. In the advanced setting, time until *definitive* deterioration (TUDD) with or without death may be more appropriate. The TUDD option differs from the first option, as a drop in a pre-specified threshold defines the first, whereas a TUDD is a drop below the pre-defined threshold where HRQoL does not rebound to a value above this threshold. Other definitions can also be applied. For example, rather than using baseline as the first reference measurement, the best HRQoL score could also be used. For a full review of options, see Anota et al. (2016) [66].

The TUD methodology is familiar within the clinical setting, as it relies on the estimation of Kaplan-Meier curves, and Cox proportional hazards models are used to estimate hazard ratios with 95% confidence intervals. Aside from ensuring that the definition of TUD is appropriate for a study, other considerations include treating time as an interval, rather than as being continuous. That is, the HRQoL drop could have occurred at any point between the assessments of HRQoL (unless diary data is being collected) and it is important to account for this in the analyses. Other thresholds may also be tested as sensitivity analyses, especially when a well-established clinically meaningful threshold is not available.

Fiteni et al. (2016) present an example of applying TUDD to provide further context for results [69]. In their study, HRQoL data came from a trial of adults aged 70–89 years with advanced non-small cell lung cancer. Data were analyzed after the primary analysis had identified a survival benefit for the paclitaxel doublet regimen compared with single-agent chemotherapy. Grade 3–4 AEs were identified in the doublet arm. Given these findings, the authors sought to compare HRQoL between the treatment arms. The EORTC QLQ-C30 was used, along with a pre-specified minimal clinically important difference of ≥5 points. Overall, the authors found, for the EORTC QLQ-C30 domains of Physical Functioning and Nausea and Vomiting, that time until a ≥5 point TUDD was significantly longer in the doublet arm, and that for the remaining domains there was no significant difference in TUDD between the arms. These results indicated that, despite increased toxicity in the doublet arm, patients' HRQoL was not adversely impacted.

20.6.2.3 Q-TWiST

Quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis was designed to combine measures of survival intervals and HRQoL, in order to estimate and compare the overall effects of different treatments [70]. In the Q-TWiST analysis, total survival is partitioned into three time periods: (1) time before disease progression without AE grade ≥3 toxicity (TWiST), (2) time with AE grade ≥3 toxicity (TOX), and (3) time after disease progression (REL). If TWiST is equal to perfect health (i.e., equal to 1) and a day of treatment TOX is worth half of that, such that the quality of 2 days of experiencing toxicity is equal to 1 day of perfect health, then the TOX coefficient would be equal to 0.5. This model attempts to balance the impacts of toxicity, OS, and HRQoL. It is written as an equation:

$$\text{Q-TWiST} = (\text{uTOX} \times \text{TOX}) + (\text{uTWiST} \times \text{TWiST}) + (\text{uREL} + \text{REL})$$

where uTWIST is the utility for the time without symptoms of disease or grade 3 or 4 toxicity due to treatment, uTOX is the utility for the time with a grade 3 or 4 toxicity, and uREL is the utility for the time after tumor progression. Mean Q-TWiST is calculated by multiplying times spent in each health state by their respective utility. This method provides an integration of both clinical outcomes (i.e., OS) and subjective PRO outcomes (i.e., HRQoL).

Satoh et al.(2014) [71] used Q-TWiST to investigate the impact of chemotherapy either alone or combined with the administration of trastuzumab in gastric cancer patients. The authors used a TWiST utility of 1.0; 0.58 was selected for REL based on published literature, and TOX was set to the same value as REL, a conservative estimate. The authors concluded that the addition of trastuzumab improved OS and did not negatively impact patients' HRQoL and that quality-adjusted survival was longer in the trastuzumab arm. In addition to the Q-TWiST analysis, a TUDD analysis was conducted and it was found that patients in the trastuzumab arm had a longer TUD in their HRQoL scores.

Currently there are limited guidelines for determining what constitutes a meaningful difference between treatment arms for Q-TWiST analyses. Revicki et al. (2006) [72] published initial recommendations for what might be a clinically important difference based on the literature. Their recommendation for a Q-TWiST of 10% for OS in a study has not been rigorously investigated and was presented by the authors with a large number of caveats [72]. In addition, the typical Q-TWiST model does not include a term for interaction between TOX and TWiST. The inclusion of such a term would allow for the consideration of, say, 2 months of TOX with 2 months of disease-free survival compared with 2 months of TOX and 2 years of disease-free survival. However, the Q-TWiST model does improve upon endpoints such as OS with the inclusion of HRQoL and toxicity.

20.7 Interpretation of Results

20.7.1 Presenting PRO Measures to Patients and Clinicians

Incorporating PRO measures into clinical practice has yet to fully take off, owing to the perceived barriers of implementing such a practice. Such barriers may include difficulty in interpreting scores, choosing an appropriate PRO measure, time, and burden. From the patient's perspective, the importance of both high-quality technical care and high-quality patient-centered care is essential. To quote from an article published by a cancer survivor: "Clearly, to maximize patients' quantity and quality of life, care delivery organizations need to emphasize both the "medical" and "care" aspects of medical care" [73].

There are four dimensions for the inclusion of PRO measures in clinical practice. The first two focus on data aggregation; in other words, considering the data at the individual or the group level. The other two dimensions focus on how the data will be used; that is, either directly with the patient in doctor-patient consultations or during multidisciplinary team meetings for patient management [74]. In this section, we will focus on information that is shared between patients and clinicians.

PRO data can be used either as part of a clinical decision-making process, where patients are shown the results of studies investigating the treatment under consideration so as to weigh the risks and benefits of a treatment, or, alternatively, to monitor their own progress, and thus inform patient management. Past research has shown that patients correctly interpret group mean HRQoL scores 85–95% of the time [75]. Further building on this work, Snyder et al. presented a series of graphs depicting progression of HRQoL and symptoms to patients, clinicians, and researchers to determine the ideal formatting for presenting this information. The authors concluded that, in general, higher = better for the directionality of line graphs and that the inclusion of a threshold line to differentiate between normal and scores of concern aided in the interpretation of PRO information [76].

20.7.2 Minimally Important Differences (MIDs)

An MID is defined as the smallest difference that a patient perceives as beneficial or harmful and that may lead a physician to alter the patient's care [77]. Numerous attempts have been made to establish a methodology for defining how to interpret scores from PRO measures in order to contextualize the impact of the change. Since the publication of the landmark paper that first proposed establishing and using minimal clinically important differences (MCIDs), other terminologies have been proposed, such as clinically important differences, minimally detectable differences, and subjectively significant differences. Some of the more recent definitions have made only small tweaks to the original definition and others have created more distinct definitions, and this continues to lead to some confusion in the literature (see Table 20.2, adapted from King (2011) [79]). In this chapter, we focus on definitions that attempt to establish differences that are meaningful to patients *and* clinicians, and we use the term MID.

Estimating an MID also allows us to understand whether a change observed over time or between groups is meaningful. For example, an MID would tell us whether a 5-point improvement on, say, the Physical Functioning domain on the SF-36 would be considered important or significant to a particular patient population [89]. In studies with large sample sizes, we may observe differences between two groups (or over time) that are considered statistically significant at the 0.05 level, but without considering the MID, we cannot conclude whether the differences are truly important or impactful for either patients or clinicians [90]. Using an MID to guide the interpretation of results can lead us to draw more meaningful conclusions, rather than relying on statistical significance. MIDs play important roles in

Table 20.2 Evolution of key terms and definitions related to the minimal important difference

Study (year)	Term	Definition	Method used and/or key distinction
Guyatt et al. (1987) [78]	Minimal clinically important difference (MCID)	MCID not defined, but used as a definition of responsiveness: 'the ability of evaluative instruments to detect minimal clinically important differences'	Change induced by an intervention of known efficacy
Jaeschke et al. (1989) [77]	Minimal clinically important difference (MCID)	The smallest difference that patients perceive as beneficial and that would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management	Global transition item ('how much has your <domain of HRQoL> changed in the past <time period>'), with the threshold based on the change in HRQoL (measured prospectively) in patients who report minimal change (on the global transition item), either for better or for worse
Osoba et al. (1998) [68]	Subjective significant difference (SDD)	The smallest change, either beneficial or deleterious, that is perceptible (discernable) to the subject	As per Jaeschke et al. [77], the important distinction is in the definition: meaningfulness is based entirely on the patient's self-assessment of the magnitude of change (note that 'perceptible (discernable)' is similar to 'detectable' in Norman et al. [31]
Guyatt et al. (2002) [79]	Minimal important difference (MID)	The smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management	Methodology is not strictly prescribed; authors suggest corroboration across 'anchor- and distribution-based' methods. The authors note that the MID is the threshold between trivial and small-but-important change. The authors also note that 'subjectively significant' is a conceptually congruent alternative label for 'minimally important'
Schünemann et al. (2005) [80]	Minimal important difference (MID)	The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the patient's management	Methodology is not strictly prescribed, but should be patient-based if possible (and while not specified, the definition implies those patients should be 'fully informed'). If proxies must be used, they should be instructed to focus on what they believe patients consider important (similarly, proxies should be 'fully informed')

(continued)

Table 20.2 (continued)

Study (year)	Term	Definition	Method used and/or key distinction
Sloan et al. (2002) [16]	Clinical significance	Goes beyond statistical significance to identify whether the statistically significant difference is large enough to have implications for patient care	Anchor- and distribution-based methods as described by Guyatt et al. [79] (the methods paper from the Clinical Significance Consensus Meeting Group of the Symposium on the Clinical Significance of Quality-of-Life Measures in Cancer Patients, Mayo Clinic [Rochester, MN, USA])
Norman et al. (2003) [31]	Clinically important differences (CIDs)	Differences that are clinically important (as determined by the method of quantification), but are not necessarily in any sense minimal	Anchor-based method involving longitudinal follow-up to determine whether subgroups can be identified that have clinically different outcomes, such as re-hospitalization, relapse of cancer, Medical Research Council grading, or different interventions
Wyrwich et al. (2005) [81]	Clinically significant change	A difference score that is large enough to have an implication for the patient's treatment or care; sometimes corresponds to what a patient might recognize as an MID	Anchor- and distribution-based methods as described by Guyatt et al. [79]
De Vet et al. (2006) [82]	Minimally important change (MIC)	A change that patients would consider important to reach in their situation, dependent on baseline values or severity of disease, on the type of intervention, and on the duration of the follow-up period	Anchor-based methods are preferred, as they include a definition of what is minimally important
Norman et al. (2003) [31]	Minimally detectable difference (MDD)	As per Jaeschke et al. [77]—same definition, different term	As per Jaeschke et al. [77]. The important distinction is in the terminology: 'clinically important' is dropped in favor of 'detectable' to more accurately reflect the quantification method (i.e., patients who report minimal change on the global transition item)
Wyrwich et al. (1999) [83]	Standard error of measurement (SEM)	The standard error in an observed score that obscures the true score	$\text{SEM} = \text{SD}\sqrt{(1-r)}$, where SD = standard deviation of the sample and r = reliability of the scale A theoretically fixed psychometric property of an instrument or scale that takes into consideration the possibility that some of the observed change may be due to random measurement error

Table 20.2 (continued)

Study (year)	Term	Definition	Method used and/or key distinction
Beaton et al. (2001) [84] and De Vet et al. (2006) [85]	Minimum detectable change (MDC)	Minimum change (at an individual level) detectable given the measurement error of the instrument (or scale)	MDC (95% confidence level) = 1.96 × $\sqrt{2}$ × SEM, where SEM is as above; 1.96 is derived from the 95% confidence interval of no change and $\sqrt{2}$ is included because two measurements are involved in measuring change (e.g., before and after an intervention or clinically significant event)
Beckerman et al. (2001) [86]	Smallest real difference (SRD)	The smallest measurement change that can be interpreted as a real difference (i.e., beyond zero), considering chance variation or measurement error	SRD = 1.96 × $\sqrt{2}$ × SEM (= MDC above)
Angst et al. (2001) [87]	Smallest statistically detectable difference (SDD)	The smallest mean change over time (within a group) that is statistically significantly different from zero	For a given sample size of n (number of patients for whom change is measured), the two-sided type-I error rate (α) and power ($1-\beta$, where β = one-sided type-II error rate) are shown by: SDD = SD ($z\alpha + z\beta$)/ $\sqrt{(n/2)}$, where $z\alpha$ and $z\beta$ are the values of the standard normal distribution (mean = 0, SD = 1) for α and β, respectively

HRQoL Health-related quality of life
SD Standard deviation
This table is adapted from King (2011) [88]

patient-centered care and shared decision-making, as they incorporate both patient and physician perspectives [60].

According to Cella et al. (2002), improvements in HRQoL are not equivalent in their impact to declines. That is, patients perceive a small improvement in HRQoL as potentially *more* impactful than a large decline [91]. This suggests that we should not consider MID improvements and deteriorations as equivalent when considering the effects of a treatment or management strategy on a patient's HRQoL [77]. For example, in a study examining improvements and deteriorations in the EORTC QLC-C30, Cocks et al. (2012) showed that the MID threshold changed by domain (e.g., Appetite Loss vs. Social Functioning) in addition to direction (e.g., improvement or decline) [92].

There are two primary methods for estimating an MID: anchor-based and distribution-based approaches. Anchor-based approaches are the most widely used and involve "anchoring" HRQoL scores to a meaningful anchor. Anchoring can be done for specific individuals or for larger groups of patients [79]. Different types of anchor-based based approaches can be used. Sometimes an objective clinical anchor is used to establish a meaningful difference and at other times a patient-reported anchor is used. For example, a Physical Functioning score could be compared to a clinician-rated performance status measure (e.g., ECOG) or an objective 6-min walk test. Another anchor-based approach is a global transition question (i.e., asking patients to rate their health compared with that at a previous point in time) [88]. In the literature, anchor-based approaches are considered optimal, as they directly incorporate the patient or clinician perspective to determine whether a change is meaningful. However, anchor-based approaches also rely on selecting appropriate anchors that are correlated (0.30 or greater) with the PRO measure of interest. In addition, anchors depend on patient and clinician recall of how the patient felt at a previous time point (which may not be accurate and can be impacted by factors such as response shift) [79]. Furthermore, as there is a great deal of variability between individual patients and groups of patients, it is difficult to know whether an MID calculated by an anchor-based method is generalizable.

Another method for establishing an MID is a distribution-based approach. This approach can be achieved by using the *standard error of measurement* to link the reliability of the PRO instrument (e.g., Cronbach's alpha) to the standard deviation of the population. Additionally, effect sizes such Cohen's D (standardized mean difference) are commonly used to compare results across different PRO measures [88]. A systematic review of effect sizes used to calculate MIDs has led to the use of 0.50 of a standard deviation as a commonly accepted MID for many scores. In the PRO literature, 0.50 of a standard deviation is a generally accepted way to determine whether there is an important difference in scores [31]. Distribution based approaches are data-driven and do not incorporate external anchors or patient input [88].

Although there is no consensus on the best approach for estimating an MID in oncology, PRO experts have suggested that considering conclusions from *both* distribution- and anchor-based approaches offers the strongest evidence [93]. In addition, experiences from clinical trials may be used to guide and inform MID identification. Finally, MIDs can vary both by patient population (e.g., disease types, age groups) and context (e.g., community clinical, hospital, clinical trial) [93]. Thus, when determining an MID for a particular PRO measure some important considerations include (1) the study population, (2) the setting, and (3) the type of approach used to estimate the MID.

20.8 Summary

In this chapter we have touched upon many important issues for including PRO measures to represent the patient's perspective on the impact of cancer and its treatment. Because the inclusion of HRQoL and other PRO measures can provide critical clinical information that is more nuanced than OS alone, the role of these measures in any study needs to be carefully and explicitly defined. This can be achieved via the thoughtful development of an analysis plan where the goals of the study, as well as the specific aims, are clearly laid out. Many of the guidelines outlined within this chapter should not be taken to be prescriptive. Each study requires consideration of the research questions and the specific treatment or study population. With careful planning, such consideration ultimately leads to results that are interpretable and meaningful to all stakeholders, but especially to the patients being treated for cancer.

References

1. American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol. 1996;14(2):671–9. https://doi.org/10.1200/JCO.1996.14.2.671.
2. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Heal. 2007;10:S125–37. https://doi.org/10.1111/j.1524-4733.2007.00275.x.
3. Ware JJ, Sherbourne C. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30(6):473–83. https://doi.org/10.1097/00005650-199206000-00002.
4. EuroQol Group. EuroQol—a new facility for the measurement of health related quality of life. Health Policy (New York). 1990;16:199–208. https://doi.org/10.1016/0168-8510(90)90421-9.
5. EuroQol – EQ-5D-3L value sets. http://www.euroqol.org/about-eq-5d/valuation-of-eq-5d/eq-5d-3l-value-sets.html. Accessed 28 Apr 2017.
6. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365–76. https://doi.org/10.1093/JNCI/85.5.365.
7. Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. Eur J Cancer. 1994;30(5):635–42. https://doi.org/10.1016/0959-8049(94)90535-5.
8. Cella D, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy Scale: development and validation of the general measure. J Clin Oncol. 1993;11(3):570–9. http://www.ncbi.nlm.nih.gov/pubmed/8445433
9. Luckett T, King MT, Butow PN, et al. Choosing between the EORTC QLQ-C30 and FACT-G for measuring health-related quality of life in cancer clinical research: issues, evidence and recommendations. Ann Oncol. 2011;22(10):2179–90. https://doi.org/10.1093/annonc/mdq721.
10. Bjordal K, De Graeff A, Fayers PM, et al. A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H and N35) in head and neck patients. Eur J Cancer. 2000;36(14):1796–807. https://doi.org/10.1016/S0959-8049(00)00186-6.
11. Sprangers MAG, Groenvald M, Arraras JI, et al. The European Organization for Research and Treatment of Cancer Breast cancer- specific quality-of-life questionnaire module: first results from a three-country field study. J Clin Oncol. 1996;14(10):2756–68. https://doi.org/10.1200/jco.1996.14.10.2756.
12. EORTC. Why do we need modules? http://groups.eortc.be/qol/why-do-we-need-modules. Accessed 28 Apr 2017.
13. Brady MJ, Cella DF, Mo F, et al. Reliability and validity of the functional assessment of cancer therapy-breast quality-of-life instrument. J Clin Oncol. 1997;15(3):974–86.
14. FACIT. Questionnaires. http://www.facit.org/facitorg/questionnaires. Accessed 28 Apr 2017.
15. Brucker PS, Yost K, Cashy J, Webster K, Cella D. General population and cancer patient norms for the Functional Assessment of Cancer Therapy-General (FACT-G). Eval Health Prof. 2005;28(2):192–211. https://doi.org/10.1177/0163278705275341.

16. Sloan J, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C. Assessing the clinical significance of single items relative to summated scores. Mayo Clin Proc. 2002;77(5):479–87. https://doi.org/10.1016/S0149-2918(02)85090-1.
17. Basch E, Reeve B, Cleeland C, et al. Development of the patient-reported version of the common terminology criteria for adverse events (pro-CTCAE). Value Heal. 2010;13(7):A274–5. https://doi.org/10.1016/S1098-3015(11)72017-4.
18. Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center. Qual Life Res. 2010;19(5):677–85. https://doi.org/10.1007/s11136-010-9634-4.
19. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol. 2009;36:2061–6. https://doi.org/10.3899/jrheum.090358.
20. Atkinson TM, Li Y, Coffey CW, et al. Reliability of adverse symptom event reporting by clinicians. Qual Life Res. 2012;21(7):1159–64. https://doi.org/10.1007/s11136-011-0031-4.
21. Jensen RE, Potosky AL, Moinpour CM, et al. United States population-based estimates of patient-reported outcomes measurement information system symptom and functional status reference values for individuals with cancer. J Clin Oncol. 2017;35(17):1913–20. https://doi.org/10.1200/JCO.2016.71.4410.
22. Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life. The Patient-Generated Index. Med Care. 1994;32(11):1109–26.
23. Macduff C, Russell E. The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. Qual Life Res. 1998;7(8):761–9. https://doi.org/10.1023/A:1008831209706.
24. Aburub AS, Gagnon B, Rodriguez AM, Mayo NE. Using a personalized measure (Patient Generated Index (PGI)) to identify what matters to people with cancer. Support Care Cancer. 2016;24(1):437–45. https://doi.org/10.1007/s00520-015-2821-7.
25. Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials. JAMA. 2013;309(8):814. https://doi.org/10.1001/jama.2013.879.
26. Patrick D. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. Value Heal. 2013;16(4):455–6. https://doi.org/10.1016/j.jval.2013.04.001.
27. Aaronson N, Choucair A, Elliott T. User's guide to implementing patient-reported outcomes assessment in clinical practice. 2011 Jan:57. http://www.isoqol.org/UserFiles/file/UsersGuide.pdf.
28. About HealthMeasures Scores. http://www.healthmeasures.net/score-and-interpret/about-healthmeasures-scores.
29. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Heal. 2008;11(2):322–33. https://doi.org/10.1111/j.1524-4733.2007.00231.x.
30. Jensen RE, Rothrock NE, DeWitt EM, et al. The role of technical advances in the adoption and integration of patient-reported outcomes in clinical care. Med Care. 2015;53(2):153–9. https://doi.org/10.1097/MLR.0000000000000289.
31. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582–92. https://doi.org/10.1097/01.MLR.0000062554.74615.4C.
32. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharm Stat. 2004;14(1):97–110. https://doi.org/10.1081/BIP-120028508.
33. Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. J Clin Oncol. 2011;29(1):89–96. https://doi.org/10.1200/JCO.2010.28.0107.
34. Bernhard JURG, Gusset H, Rny CHU. Practical issues in quality of life assessment in multicentre trials conducted by the Swiss Group for Clinical Cancer Research. Stat Med. 1998;17:633–9.

35. Mercieca-Bebber R, Palmer MJ, Brundage M, Calvert M, Stockler MR, King MT. Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review. BMJ Open. 2016;6(6):e010938. https://doi.org/10.1136/bmjopen-2015-010938.
36. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70. http://www.citeulike.org/user/santi515/article/4294367.
37. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2. https://doi.org/10.1093/biomet/75.4.800.
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1). https://doi.org/10.2307/2346101.
39. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41. https://doi.org/10.2307/2335942.
40. Reeve BB, Stover AM, Jensen RE, et al. Impact of diagnosis and treatment of clinically localized prostate cancer on health-related quality of life for older Americans: a population-based study. Cancer. 2012;118(22):5679–87. https://doi.org/10.1002/cncr.27578.
41. Keating NL, Norredam M, Landrum MB, Huskamp HA, Meara E. Physical and mental health status of older long-term cancer survivors. J Am Geriatr Soc. 2005;53(12):2145–52. https://doi.org/10.1111/j.1532-5415.2005.00507.x.
42. Wang S-Y, Hsu SH, Gross CP, et al. Association between time since cancer diagnosis and health-related quality of life: a population-level analysis. Value Heal. 2016;19(5):631–8. https://doi.org/10.1016/j.jval.2016.02.010.
43. Fromme EK, Eilers KM, Mori M, Hsieh YC, Beer TM. How accurate is clinician reporting of chemotherapy adverse effects? A comparison with patient-reported symptoms from the Quality-of-Life Questionnaire C30. J Clin Oncol. 2004;22(17):3485–90. https://doi.org/10.1200/JCO.2004.03.025.
44. Pakhomov SV, Jacobsen SJ, Chute CG, Roger VL. Agreement between patient-reported symptoms and their documentation in the medical record. Am J Manag Care. 2008;14(8):530–9. https://doi.org/10.1016/j.bbi.2008.05.010.
45. Basch E, Jia X, Heller G, et al. Adverse symptom event reporting by patients vs clinicians: Relationships with clinical outcomes. J Natl Cancer Inst. 2009;101(23):1624–32. https://doi.org/10.1093/jnci/djp386.
46. Quinten C, Maringwa J, Gotay CC, et al. Patient self-reports of symptoms and clinician ratings as predictors of overall cancer survival. J Natl Cancer Inst. 2010;103(24):1851–8. https://doi.org/10.1093/jnci/djr485.
47. Gotay CC, Kawamoto CT, Bottomley A, Efficace F. The prognostic significance of patient-reported outcomes in cancer clinical trials. J Clin Oncol. 2008;26(8):1338–45. https://doi.org/10.1200/JCO.2007.13.9337.
48. Diouf M, Filleron T, Barbare J-C, et al. The added value of quality of life (QoL) for prognosis of overall survival in patients with palliative hepatocellular carcinoma. J Hepatol. 2013;58(3):509–21. https://doi.org/10.1016/j.jhep.2012.11.019.
49. European Medicines Agency. Guideline on missing data in confirmatory clinical trials; 2008. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.
50. Fairclough D. Design and analysis of quality of life studies in clinical trials. 2nd ed. New York, NY: Taylor & Francis Group; 2010. https://books.google.com/books?hl=en&lr=&id=FV-SU8IZDcAC&oi=fnd&pg=PP1&dq=Design+and+analysis+of+quality+of+life+studies+in+clinical+trials&ots=JMzjj07Z_t&sig=Iv_r6YvABchDiqyq_3ToKxofGNc#v=onepage&q=Design+and+analysis+of+quality+of+life+stu.
51. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. Stat Med. 2003;22(5):795–809. https://doi.org/10.1002/sim.1397.
52. Raboud JM, Singer J, Thorne A, Schechter MT, Shafran SD. Estimating the effect of treatment on quality of life in the presence of missing data due to drop-out and death. Qual Life Res. 1998;7(6):487–94. https://doi.org/10.1023/A:1008870223350.

53. Fairclough DL, Fetting JH, Cella D, Wonson W, Moinpour CM. Quality of life and quality adjusted survival for breast cancer patients receiving adjuvant therapy. Qual Life Res. 1999;8(8):723–31. https://doi.org/10.1023/A:1008806828316.
54. Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. Stat Sci. 2009;24(2):211–22. https://doi.org/10.1214/09-STS293.
55. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Stat Med. 1992;11(14–15):1861–70. http://www.ncbi.nlm.nih.gov/pubmed/1480878.
56. Ribaudo HJ, Thompson SG, Allen-Mersh TG. A joint analysis of quality of life and survival using a random effect selection model. Stat Med. 2000;19(23):3237–50. http://www.ncbi.nlm.nih.gov/pubmed/11113957.
57. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. Stat Med. 1999;18(10):1215–33. http://www.ncbi.nlm.nih.gov/pubmed/10363341.
58. Vonesh EF, Greene T, Schluchter MD. Shared parameter models for the joint analysis of longitudinal data and event times. Stat Med. 2006;25(1):143–63. https://doi.org/10.1002/sim.2249.
59. Chi Y-Y, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. Biometrics. 2006;62(2):432–45. https://doi.org/10.1111/j.1541-0420.2005.00448.x.
60. Elashoff RM, Li G, Li N. An approach to joint analysis of longitudinal measurements and competing risks failure time data. Stat Med. 2007;26(14):2813–35. https://doi.org/10.1002/sim.2749.
61. Law NJ, Taylor JMG, Sandler H. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. Biostatistics. 2002;3(4):547–63. https://doi.org/10.1093/biostatistics/3.4.547.
62. Yu M, Law NJ, Taylor JMG, Sandler HM. Joint longitudinal-survival-cure models and their application to prostate cancer. Stat Sin. 2004;14:835–62. http://www3.stat.sinica.edu.tw/statistica/oldpdf/A14n310.pdf.
63. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol. 1977;39(1):1–38. https://doi.org/10.2307/2984875.
64. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. Biometrics. 1986;42(4):805–20. http://www.ncbi.nlm.nih.gov/pubmed/3814725.
65. Siddiqui O, Hung HMJ, O'Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. J Biopharm Stat. 2009;19(2):227–46. https://doi.org/10.1080/10543400802609797.
66. Anota A, Hamidou Z, Paget-Bailly S, et al. Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Qual Life Res. 2015;24(1):5–18. https://doi.org/10.1007/s11136-013-0583-6.
67. Bonnetain F, Dahan L, Maillard E, et al. Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. Eur J Cancer. 2010;46(15):2753–62. https://doi.org/10.1016/j.ejca.2010.07.023.
68. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol. 1998;16(1):139–44. https://doi.org/10.1200/JCO.1998.16.1.139.
69. Fiteni F, Anota A, Bonnetain F, et al. Health-related quality of life in elderly patients with advanced non-small cell lung cancer comparing carboplatin and weekly paclitaxel doublet chemotherapy with monotherapy. Eur Respir J. 2016;48(3):861–72. https://doi.org/10.1183/13993003.01695-2015.
70. Gelber RD, Goldhirsch A. A new endpoint for the assessment of adjuvant therapy in postmenopausal women with operable breast cancer. J Clin Oncol. 1986;4(12):1772–9. https://doi.org/10.1200/JCO.1986.4.12.1772.
71. Satoh T, Bang YJ, Gotovkin EA, Hamamoto Y, Kang YK, Moiseyenko VM, et al. Quality of life in the trastuzumab for gastric cancer trial. Oncologist. 2014;19(7):712–9.

72. Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. Qual Life Res. 2006;15(3):411–23. https://doi.org/10.1007/s11136-005-1579-7.
73. Arora NK. Importance of patient-centered care in enhancing patient well-being: a cancer survivor's perspective. Qual Life Res. 2009;18(1):1–4. https://doi.org/10.1007/s11136-008-9415-5.
74. Greenhalgh J. The applications of PROs in clinical practice: what are they, do they work, and why? Qual Life Res. 2009;18(1):115–23. https://doi.org/10.1007/s11136-008-9430-6.
75. Brundage M, Feldman-Stewart D, Leis A, et al. Communicating quality of life information to cancer patients: a study of six presentation formats. J Clin Oncol. 2005;23(28):6949–56. https://doi.org/10.1200/JCO.2005.12.514.
76. Snyder CF, Smith KC, Bantug ET, et al. What do these scores mean? Presenting patient-reported outcomes data to patients and clinicians to improve interpretability. Cancer. 2017;123(10):1848–59. https://doi.org/10.1002/cncr.30530.
77. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15. http://www.ncbi.nlm.nih.gov/pubmed/2691207.
78. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8. http://www.ncbi.nlm.nih.gov/pubmed/3818871.
79. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83. https://doi.org/10.1016/S0025-6196(11)61793-X.
80. Schünemann HJ, Guyatt GH. Commentary—goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005;40(2):593–7. https://doi.org/10.1111/j.1475-6773.2005.00374.x.
81. Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. Qual Life Res. 2005;14(2):285–95. http://www.ncbi.nlm.nih.gov/pubmed/15892420.
82. de Vet HCW, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. J Rheumatol. 2006;33(2):434. Author reply 435. http://www.ncbi.nlm.nih.gov/pubmed/16465677.
83. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999;52(9):861–73. http://www.ncbi.nlm.nih.gov/pubmed/10529027.
84. Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome measures in rheumatology. Minimal clinically important difference. J Rheumatol. 2001;28(2):400–5. http://www.ncbi.nlm.nih.gov/pubmed/11246687.
85. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4(1):54. https://doi.org/10.1186/1477-7525-4-54.
86. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res. 2001;10(7):571–8. http://www.ncbi.nlm.nih.gov/pubmed/11822790.
87. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum. 2001;45(4):384–91. https://doi.org/10.1002/1529-0131(200108)45:4<384::AID-ART352>3.0.CO;2-0.
88. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res. 2011;11(2):171–84. https://doi.org/10.1586/erp.11.9.

89. Johnston BC, Ebrahim S, Carrasco-Labra A, et al. Minimally important difference estimates and methods: a protocol. BMJ Open. 2015;5(10):e007953. https://doi.org/10.1136/bmjopen-2015-007953.
90. Eton DT, Cella D, Yost KJ, et al. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004;57(9):898–910. https://doi.org/10.1016/j.jclinepi.2004.01.012.
91. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res. 2002;11(3):207–21. https://doi.org/10.1023/A:1015276414526.
92. Cocks K, King MT, Velikova G, De Castro G, St-James MM, Fayers PM, Brown JM. Evidence-based guidelines for interpreting change scores for the European organisation for the research and treatment of cancer quality of life questionnaire core 30. Eur J Cancer. 2012;48(11):1713–21.
93. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9. https://doi.org/10.1016/j.jclinepi.2007.03.012.

Index

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3