Practice online to succeed in real life:
Get research exercises at MediaCenter.Thieme.com!

Simply visit MediaCenter.Thieme.com and, when prompted during the registration process, enter the code below to get started today.

85RH-9R22-N84X-MA3N

<table>
<thead>
<tr>
<th>WINDOWS &amp; MAC</th>
<th>TABLET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended Browser(s)</strong></td>
<td></td>
</tr>
<tr>
<td>Recent browser versions on all major platforms and any mobile operating system that supports HTML5 video playback. All browsers should have JavaScript enabled.</td>
<td></td>
</tr>
<tr>
<td><strong>Flash Player Plug-in</strong></td>
<td></td>
</tr>
<tr>
<td>Flash Player 9 or higher. For Mac users, ATI Rage 128 GPU doesn’t support full-screen mode with hardware scaling.</td>
<td>Tablet PCs with Android OS support Flash 10.1.</td>
</tr>
<tr>
<td><strong>Recommended for optimal usage experience</strong></td>
<td></td>
</tr>
<tr>
<td>Monitor resolutions:</td>
<td></td>
</tr>
<tr>
<td>• Normal (4:3) 1024×768 or higher</td>
<td>WiFi or cellular data connection is required.</td>
</tr>
<tr>
<td>• Widescreen (16:9) 1280×720 or higher</td>
<td></td>
</tr>
<tr>
<td>• Widescreen (16:10) 1440×900 or higher</td>
<td></td>
</tr>
<tr>
<td>A high-speed internet connection (minimum 384 Kps) is suggested.</td>
<td></td>
</tr>
</tbody>
</table>
Research Methods in Radiology
A Practical Guide

Andrea S. Doria, MD, PhD, MSc
Professor of Radiology and Associate Chair of Research (Injury, Repair and Inflammation)
Department of Medical Imaging
University of Toronto
Research Director, Radiologist,
Department of Diagnostic Imaging
Senior Scientist, Translational Medicine, Research Institute
The Hospital for Sick Children
Toronto, Ontario, Canada

George Tomlinson, MSc, BSc
Associate Professor
Institute of Health Policy, Management and Evaluation
Dalla Lana School of Public Health
Departments of Medicine and Medical Imaging
University of Toronto
Scientist
Department of Medicine
University Health Network
Toronto, Ontario, Canada

Joseph Beyene, PhD, MSc, BSc
Professor
John D. Cameron Endowed Chair in Genetic Epidemiology
Department of Health Research Methods, Evidence, and Impact
McMaster University
Hamilton, Ontario, Canada

Rahim Moineddin, PhD
Professor
Department of Family and Community Medicine
University of Toronto
Toronto, Ontario, Canada

Associate Editors:
Christopher Meaney, MSc
Biostatistician
Department of Family and Community Medicine
PhD candidate
Department of Biostatistics
University of Toronto
Toronto, Ontario, Canada

Mateen Shaikh, PhD
Post-Doctoral Fellow
Department of Clinical Epidemiology & Biostatistics (Faculty of Health Sciences)
McMaster University
Hamilton, Ontario, Canada

112 illustrations

Thieme
New York • Stuttgart • Delhi • Rio de Janeiro
Important note: Medicine is an ever-changing science undergoing continual development. Research and clinical experience are continually expanding our knowledge, in particular our knowledge of proper treatment and drug therapy. Insofar as this book mentions any dosage or application, readers may rest assured that the authors, editors, and publishers have made every effort to ensure that such references are in accordance with the state of knowledge at the time of production of the book.

Nevertheless, this does not involve, imply, or express any guarantee or responsibility on the part of the publishers in respect to any dosage instructions and forms of applications stated in the book. Every user is requested to examine carefully the manufacturers' leaflets accompanying each drug and to check, if necessary in consultation with a physician or specialist, whether the dosage schedules mentioned therein or the contraindications stated by the manufacturers differ from the statements made in the present book. Such examination is particularly important with drugs that are either rarely used or have been newly released on the market. Every dosage schedule or every form of application used is entirely at the user's own risk and responsibility. The authors and publishers request every user to report to the publishers any discrepancies or inaccuracies noticed. If errors in this work are found after publication, errata will be posted at www.thieme.com on the product description page.

Some of the product names, patents, and registered designs referred to in this book are in fact registered trademarks or proprietary names even though specific reference to this fact is not always made in the text. Therefore, the appearance of a name without designation as proprietary is not to be construed as a representation by the publisher that it is in the public domain.
To my father who has been my great mentor, advisor, and friend throughout my life.

To my mother who taught me beautiful lessons of love during the short period of time that I had the privileged to live by her side.

To my husband, son, stepson, and extended family accrued throughout the pathway of life, who have taught me values of loyalty, honesty, companionship, and true love.

To academic leaders such as Drs. Paul Babyn, Walter Kucharczyk, and Brian Feldman who believed in me and gave me the opportunity and support to pursue my burning desire of learning research methods as a young radiologist.

To my patients whose journeys and battles have been the inspiration of my life, and the main reason for my life mission of trying to improve their lives through research of unanswered clinical questions.

Andrea S. Doria

The Editor and Associate Editors would also like to acknowledge the support of Kieran Murphy for the initiation of this book; the artistic work of Dr. Rayan Ahyad, radiologist and illustrator, for the artwork on the cover; and Mr. Danny Aguilar, medical graphic artist and illustrator, for other illustrations in the book.
Contents

Editors ................................................................. ix
Foreword .............................................................. xi
Preface ................................................................. xiii
Contributors ............................................................ xv

Chapter 1  Overview of Research Designs Applied to Radiology
Andreas Roposch and Andrea S. Doria ......................... 1

Chapter 2  Descriptive Statistics
Mateen Shaikh and Hadas Moshonov ........................... 13

Chapter 3  Diagnostic Tests
George Tomlinson, Gerald Lebovic, Connie Marras,
and Andrea S. Doria ................................................ 30

Chapter 4  Measurement: Validity, Reliability, and Responsiveness
Alexandra M. Easson, George Tomlinson, and
Andrea S. Doria ...................................................... 54

Chapter 5  Observational Designs
Marinka Twilt, Ivan R. Diamond, and Susanne M. Benseler .... 87

Chapter 6  Randomized Controlled Trials
Ivan R. Diamond and Brian M. Feldman ......................... 104

Chapter 7  Systematic Reviews, Evidence-Based Imaging, and
Knowledge Translation
Andrea S. Doria, Jennifer Stinson, and Prakeshkumar Shah .... 116

Chapter 8  Decision Analysis
Ava A. John-Baptiste and Sapna Rawal ........................ 138

Chapter 9  Costs and Consequences: Economic Evaluation in Radiology
Wendy J. Ungar and Richard M. Zur ............................ 153

Chapter 10 Conducting and Publishing Research
Andrea S. Doria, Peter J. Strouse, and
Thomasin Adams-Webber .......................................... 174
## Contents

**Chapter 11**  Obtaining Ethics Approval, Data Management, and Budget Formulation  
*Ravi Menezes, Yasser Karimzad, and Tuula Kalliomäki* ..........196

**Chapter 12**  Steps for Using Research as a Success Tool in Academia  
*Aine Kelly and Suresh K. Mukherji* .............................210

**Chapter 13**  Statistical Inference: Point Estimation, Confidence Intervals, and Hypothesis Testing  
*Christopher Meaney, Rahim Moineddin, and George Tomlinson* ............................................225

**Chapter 14**  Linear and Logistic Regression  
*Christopher Meaney, Mateen Shaikh, Andrea S. Doria, and Rahim Moineddin* .....................................245

**Chapter 15**  Sample Size Estimation  
*Mateen Shaikh, Joseph Beyene, and Andrea S. Doria* ...............257

**Chapter 16**  Meta-analysis  
*Joseph Beyene, Binod Neupane, and Zelalem F. Negeri* .............269

**Appendix:** A Brief Introduction to ‘R’ ............................................285

**Index** ..............................................................................................................291
Editors

Andrea S. Doria, MD, PhD, MSc, FACR

George Tomlinson, MSc, BSc

Joseph Beyene, PhD, MSc, BSc

Rahim Moineddin, PhD

Christopher Meaney, MSc

Mateen Shaikh, PhD
As both a clinical and laboratory-based researcher in pediatric radiology, I was fortunate to have had access to excellent mentoring in research design, data analysis, and strategies for publishing my work. Yet, learning how to navigate the long and sometimes arduous path from research idea to publication can be a daunting challenge to the individual beginning her/his research career.

As the discipline of radiology has transitioned from anatomic to functional imaging, and from qualitative to quantitative analysis, there is an increasingly important need for all radiologists to understand the strengths and weaknesses of our techniques along an efficacy continuum. We as radiologists must justify the use of imaging studies based on the impact on patient diagnosis, individual patient outcomes, and cost-effectiveness. In addition, we are also frequently asked to weigh in on societal outcomes of our work. Nothing has brought this area more keenly into focus as the changing guidelines for screening mammography for women with “standard risk” for breast cancer. As a result, we find ourselves in the center of a highly charged discussion on the risks and benefits of mammography. We cannot meaningfully participate in this discussion without the ability to analyze and understand the literature, research design, and statistical analysis of the data used to make these decisions.

This brings me to the strong need for a tome such as this. A recent electronic search for books on statistical methods revealed that there are books available on statistical methods related to a number of specialized fields, including methods for the social sciences, hydrology, food security, business, plant variety evaluation, and diagnostic medicine. Each deals with specific problems inherent in that particular field. Thus, trying to extrapolate the effective use of research and statistical methods for use in imaging can be quite a challenge.

The idea for Research Methods in Radiology began in 2006 as an innovative course entitled “Introduction to Research in Radiology” at the University of Toronto. In 2008, the course became a part of the curriculum of residency in radiology at the University. This book makes the main aspects of this very successful course available to a wider audience of radiologists under a single cover. Dr. Doria and co-editors Drs. Tomlinson, Beyene, and Moineddin have assembled an unparalleled team of contributors with specialized expertise in various areas of the research and publication process. The book is organized into chapters on research designs and statistical methods. Chapters on research designs discuss specific types of research commonly used in imaging studies, with detailed discussions about research design, the complex process of developing a
budget, navigating a research ethics board, translating a research design into a finished product for publication, and, finally, strategies for academic development and funding.

Chapters on statistical methods explain the range of statistical methods used in our specialty in a readable and understandable format. Both categories of chapters expertly cover the concepts underlying a specific design or statistical technique, along with their strengths, limitations, and pitfalls in minimizing bias and maximizing validity.

This book is a tremendous resource to anyone interested in imaging research, and in evaluating the imaging literature. The end result of reading this book will be not to be able to distinguish between lies and statistics, but to understand the difference between something that is true and the hope that something might be true.

George A. Taylor, MD, FACR
John A. Kirkpatrick Professor of Radiology (Pediatrics)
Harvard Medical School
Radiologist-in-Chief Emeritus
Boston Children’s Hospital
Boston, Massachusetts
Preface

Shortfalls in the quality of research studies produced by the radiology community stand in sharp contrast with emerging technologies and the need for new knowledge to address several issues in the radiological sciences. It has been shown that engaging trainees and retaining their interest throughout their training influences their decision on choosing (or not) a career in academic medicine. For this reason our group has developed and conducted a research training program targeted to radiology residents and fellows in the format of a workshop from 2006 to present. This program was initially launched within the Department of Medical Imaging of the University of Toronto, which has the largest radiology training program (residency and fellowship) in Canada. This book has been conceived on the basis of the feedback received from the workshops’ participants and as a way of improving the effectiveness of conducting research and the quality of research projects conducted by residents, fellows, radiologists, statisticians, and epidemiologists with an interest in radiology research in different parts of the world.

Although previous books have addressed different research topics in radiology, to our knowledge no prior book has showed in a simple way how to design a clinical or experimental study in radiology, providing pertinent examples (case studies) with imaging figures, graphs, and tables. In addition, this book contains exercises related to the topic under discussion in each chapter (as supplementary material) which should enable trainees to prepare themselves for Radiology Board Examinations and academic radiologists to apply the knowledge obtained in the book into the conduct of their own research studies. Further, the online ancillary component of this book provides a glossary for methodological and statistical terms, a tutorial and codes for the use of “R,” a web-free software for statistical analysis, with examples for sample size calculation, plot of receiver-operating characteristic curves, and statistical tests for different research questions.

The chapters of this book are organized into basic concepts of statistics applied to radiology, research designs (from the most to the least commonly used in the radiological sciences, including advanced designs such as decision-making modeling and economics evaluations), evidence-based imaging, data administration, types of statistical methods including regression analysis and meta-analysis, and sample size estimation for different outcomes.

Each chapter comprises information about learning objectives, basic concepts, highlights, and supplementary material (displayed in the book website only, including exercises, supplementary tables, formulas, frequently asked questions and “R” codes used in the body of text of statistical chapters).

The chapters were drafted by radiologists, epidemiologists, clinicians, statisticians, and librarians who have wide experience in multidisciplinary research. Given the wide amount
of information on research methodology and statistics available in the literature, this book was not envisioned to be a comprehensive reference book, but rather to serve as a high-yield review of research methodology for radiology trainees and a focused reference and aid tool for clinical and experimental research.

We wish you all the best in your future academic endeavors, hoping that your future research may improve the way radiologists practice medicine through high methodological standards and strong evidence basis. We sincerely hope that this book can help you achieve this final goal.
Contributors

**Thomasin Adams-Webber, MLS**  
Manager, Library and Archives Services  
Hospital Library and Archives  
The Hospital for Sick Children  
Toronto, Ontario, Canada

**Ava A. John-Baptiste, PhD**  
Assistant Professor, Anesthesia & Perioperative Medicine  
Epidemiology & Biostatistics Interfaculty Program in Public Health  
Western University  
London, Ontario, Canada

**Susanne M. Benseler MD, MSc, PhD**  
Professor of Pediatrics  
University of Calgary  
Pediatric Rheumatologist and Section Chief Pediatric Rheumatology  
Alberta Children’s Hospital  
Calgary, Alberta, Canada

**Joseph Beyene, PhD, MSc, BSc**  
Professor  
John D. Cameron Endowed Chair in Genetic Epidemiology  
Department of Health Research Methods, Evidence, and Impact  
McMaster University  
Hamilton, Ontario, Canada

**Ivan R. Diamond, MD, PhD, FRCPC**  
Radiologist  
Trillium Health Partners  
Mississauga, Ontario, Canada

**Andrea S. Doria, MD, PhD, MSc**  
Professor of Radiology and Associate Chair of Research (Injury, Repair and Inflammation)  
Department of Medical Imaging  
University of Toronto  
Research Director, Radiologist, Department of Diagnostic Imaging  
Senior Scientist, Translational Medicine, Research Institute  
The Hospital for Sick Children  
Toronto, Ontario, Canada

**Alexandra M. Easson, MSc, MD, FRCSC, FACS**  
Cancer Clinical Research Unit  
Princess Margaret Hospital  
Associate Professor  
Department of Surgery  
University of Toronto  
Toronto Ontario, Canada

**Brian M. Feldman, MD, MSc, FRCPC**  
Professor, Faculty of Medicine, and the Dalla Lana School of Public Health, University of Toronto  
Division Head, Rheumatology  
Senior Scientist, Child Health Evaluative Sciences, Research Institute  
The Hospital for Sick Children  
Toronto, Ontario, Canada

**Tuula Kalliomaki, PhD, MBA**  
Assistant Professor  
Manager, R&D, Joint Department of Medical Imaging  
Department of Medical Imaging  
University of Toronto  
Toronto, Ontario, Canada

**Ava A. John-Baptiste, PhD**  
Assistant Professor  
Anesthesia & Perioperative Medicine  
Epidemiology & Biostatistics Interfaculty Program in Public Health  
Western University  
London, Ontario, Canada

**Joseph Beyene, PhD, MSc, BSc**  
Professor  
John D. Cameron Endowed Chair in Genetic Epidemiology  
Department of Health Research Methods, Evidence, and Impact  
McMaster University  
Hamilton, Ontario, Canada

**Ivan R. Diamond, MD, PhD, FRCPC**  
Radiologist  
Trillium Health Partners  
Mississauga, Ontario, Canada

**Andrea S. Doria, MD, PhD, MSc**  
Professor of Radiology and Associate Chair of Research (Injury, Repair and Inflammation)  
Department of Medical Imaging  
University of Toronto  
Research Director, Radiologist, Department of Diagnostic Imaging  
Senior Scientist, Translational Medicine, Research Institute  
The Hospital for Sick Children  
Toronto, Ontario, Canada

**Alexandra M. Easson, MSc, MD, FRCSC, FACS**  
Cancer Clinical Research Unit  
Princess Margaret Hospital  
Associate Professor  
Department of Surgery  
University of Toronto  
Toronto Ontario, Canada

**Brian M. Feldman, MD, MSc, FRCPC**  
Professor, Faculty of Medicine, and the Dalla Lana School of Public Health, University of Toronto  
Division Head, Rheumatology  
Senior Scientist, Child Health Evaluative Sciences, Research Institute  
The Hospital for Sick Children  
Toronto, Ontario, Canada

**Tuula Kalliomaki, PhD, MBA**  
Assistant Professor  
Manager, R&D, Joint Department of Medical Imaging  
Department of Medical Imaging  
University of Toronto  
Toronto, Ontario, Canada
Yasser Karimzad, HBSc, CCRP  
Clinical Research Coordinator  
Joint Department of Medical Imaging (MSH/UHN/WCH)  
Toronto, Ontario, Canada

Aine Kelly, MD  
Professor  
Department of Radiology  
Division of Cardiothoracic Radiology  
University of Michigan Health System  
Ann Arbor, Michigan

Rahim Moineddin, PhD  
Professor  
Department of Family and Community Medicine  
University of Toronto  
Toronto, Ontario, Canada

Hadas Moshonov, PhD  
Course Instructor and Coordinator  
University of Toronto  
Toronto, Ontario, Canada

Suresh K. Mukherji, MD, MBA, FACP  
Professor & Chairman  
Walter F. Patenge Endowed Chair  
Department of Radiology  
Chief Medical Officer & Director of Health Care Delivery  
Michigan State University Health Team  
East Lansing, Michigan

Zelalem F. Negeri, MSc, BSc, PhD  
Candidate Research and Teaching Assistant  
Department of Mathematics and Statistics  
McMaster University  
Hamilton, Ontario, Canada

Binod Neupane, PhD  
Post-Doctoral Fellow  
Department of Health Research Methods, Evidence, and Impact  
Faculty of Health Sciences  
McMaster University  
Hamilton, Ontario, Canada

Sapna Rawal, MD, PhD Candidate  
Assistant Professor  
Division of Neuroradiology  
Department of Medical Imaging  
University of Toronto  
Neuroradiologist  
University Health Network  
Toronto Western Hospital  
Institute of Health Policy, Management and Evaluation  
Toronto, Ontario, Canada

Andreas Roposch, MD, MSc, FRCSEd  
Professor of Surgery and Clinical Epidemiology  
University College London  
Consultant Orthopedic Surgeon  
Great Ormond Street Hospital for Children  
London, United Kingdom
Prakeshkumar Shah, MD, MSc MBBS, MD, DCH, MRCP, FRCPCH
Professor
Departments of Pediatrics and Institute of Health Policy, Management and Evaluation
University of Toronto
CIHR Applied Research Chair in Reproductive and Child Health Services and Policy Research
Director, Canadian Neonatal Network
Mount Sinai Hospital
Toronto Ontario, Canada

Mateen Shaikh, PhD
Post-Doctoral Fellow
Department of Clinical Epidemiology & Biostatistics (Faculty of Health Sciences)
McMaster University
Hamilton, Ontario, Canada

Jennifer Stinson, RN, PhD
Professor
Lawrence S. Bloomberg Faculty of Nursing
University of Toronto
Centre for the Study of Pain
Nurse Practitioner, Senior Scientist
Child Health Evaluative Sciences, Research Institute
The Hospital for Sick Children
Toronto, Ontario, Canada

Peter J. Strouse, MD
John F. Holt Collegiate Professor of Radiology
University of Michigan
Director, Section of Pediatric Radiology
C. S. Mott Children's Hospital
Department of Radiology
University of Michigan Health System
Ann Arbor, Michigan

George Tomlinson, PhD, MSc
Associate Professor
Institute of Health Policy, Management and Evaluation
Dalla Lana School of Public Health
Departments of Medicine and Medical Imaging
University of Toronto
Scientist
Department of Medicine
University Health Network
Toronto, Ontario, Canada

Marinka Twilt MD, MSc, PhD
Assistant Professor of Pediatrics
University of Calgary
Pediatric Rheumatologist, Immunologist
Alberta Children’s Hospital
Calgary, Alberta, Canada

Wendy Ungar, MSc, PhD
Professor
Institute of Health Policy, Management and Evaluation
University of Toronto
Senior Scientist, Child Health Evaluative Sciences, Research Institute
The Hospital for Sick Children
Toronto, Ontario, Canada

Richard M. Zur, PhD
John F. Holt Collegiate Professor of Radiology
University of Michigan
Director, Section of Pediatric Radiology
C. S. Mott Children's Hospital
Department of Radiology
University of Michigan Health System
Ann Arbor, Michigan
Overview of Research Designs
Applied to Radiology
Andreas Roposch and Andrea S. Doria

Learning Objectives
- To provide an overview of available research designs in clinical research focusing on advantages and disadvantages of each type of design.
- To recognize the differences among clinical research designs.
- To identify the design of a study.
- To understand the applications of each type of research design.
- To choose an appropriate design to examine a given question.

Introduction
The fundamental goal of clinical research is to improve patient care. To do this, researchers attempt to learn the “truth” about the frequency of disease, etiology, prognosis, or therapy. There are two main categories of research design that can be used to obtain primary data and compare therapies: nonexperimental and experimental methods.

<table>
<thead>
<tr>
<th>Feature of interest</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>Who are you asking the question about?</td>
</tr>
<tr>
<td>Intervention</td>
<td>What intervention are you interested in?</td>
</tr>
<tr>
<td>Control or comparison</td>
<td>What are you comparing the intervention with?</td>
</tr>
<tr>
<td>Outcome</td>
<td>What outcome are you interested in measuring?</td>
</tr>
</tbody>
</table>

Steps of a Research Project
The initial steps in planning a study include:
- Defining the research question, which should be novel, relevant, and feasible.
- Conducting a literature search to determine whether any other similar studies are already available in the literature and to set the stage for the methodology of the current study.
- Estimating the sample size to determine the feasibility of the study in your own center (single center) or the need for a multicenter study if the number of required subjects for the study bypasses the available number of patients with a given condition in your center.
Characteristics of a good research question are (FINER):  

F: Feasible  
N: Narrow in scope: “If you hit a bell with a pebble you will get a small sound. If you hit it with a mallet you will achieve a large sound.” One of the major mistakes of young investigators in conceiving a research question is to formulate questions that are broad in scope and ambitious in number of features to be addressed.  
I: Interesting  
N: Novel  
E: Ethical  
R: Relevant  

Typically, a strong research project has one primary research question and no more than three (or four) secondary questions. A correctly formulated research question should result in a “yes/no” answer.  

Investigators can use descriptive or analytic research designs to conduct their research using primary data (Fig. 1.2). Nonexperimental (observational) and experimental research designs fall within the analytic category of designs.  

Formulating a Research Question  

Designing a proper study question to address an area of clinical uncertainty is the first step in any research initiative, and one way to focus your research is to use the PICO approach (people, patients, or population; intervention; control or comparison; and outcome). Table 1.1 shows the features of interest and corresponding questions of the PICO approach for formulating a research question.
Overview of Research Designs Applied to Radiology

available method to establish the presence or absence of the target condition. Often the reference standard is a test that, compared to the index test, is more expensive (magnetic resonance imaging [MRI], reference standard vs. ultrasound, index test), potentially more harmful (computed tomography [CT], reference standard vs. radiography, index test), more invasive (biopsy, reference standard vs. MRI, index test), or requires more logistical efforts (diagnostic expert panel).

In evaluating diagnostic tests, several study designs can be applied, depending on the stage of evaluation of a diagnostic test and the research question. The basic stage in the evaluation of a diagnostic test is concerned with test properties (Fig. 1.3). Test properties tell us how well the test can discriminate between healthy and diseased individuals. The essential property of any diagnostic test is reliability. A test is regarded as reliable if the same result is obtained when the test is repeated by the same (intra-rater or test-retest reliability) or different (inter-rater reliability) investigators (or technologies). Reliability is the prerequisite for validity, which is the degree of closeness of measurements of a quantity to that quantity’s true value.

Fig. 1.2 Types of research designs.

Nonexperimental (Observational) Research

Diagnostic Test Design

Diagnostic tests measure and classify clinical phenomena in the context of accepted knowledge on nosology. This process will include or exclude a certain disease, select a disease from a set of candidate diseases, or establish the presence of several diseases. The goal of the diagnostic process is to improve patient care and clinical outcomes.

Diagnostic tests not only provide information about the likelihood of a particular diagnosis, but also determine the next step in the management of an individual who is being tested. It is thus important that diagnostic tests are accurate. Accuracy is the ability of a test to differentiate between patients who have the condition of interest (target condition) and those who do not. The accuracy of a test is studied by comparing the results of the test under investigation (index test) with the results of a reference standard on the same participants. The reference standard (also referred to as “gold standard”) is the best
Validity in diagnostic test research tells us if the conclusion of the test is actually true. The term accuracy, typically used in diagnostic test research, encompasses both reliability and validity. An accurate test is one that has demonstrated reliability and validity. In evaluating the accuracy of a diagnostic test, the cross-sectional study design is commonly employed. Alternatives are cohort studies but also case-control studies. Once the diagnostic accuracy is established, other stages in evaluating the test should be explored to understand the relevance of the test in terms of clinical outcomes. These include understanding how a particular test affects the diagnostic workup in general; how a test affects the choice of several treatment options; and to what degree having a test improves patient outcomes such as quality of life or life expectancy. Diagnostic tests can be embedded into specific observational designs, as we discuss in the following sections of this chapter.

Cross-Sectional Design

A cross-sectional study takes place at one point (cross-section) in time. The researcher identifies a sample (study participants) from a population of interest. All study maneuvers are carried out at one point in time: disease and exposure are measured simultaneously (Fig. 1.4). This is perhaps the most commonly used design in establishing the basic properties of a diagnostic test: reliability and validity. During evaluation, we determine how well a novel diagnostic test (index test) performs by comparing its results against an established test (reference standard). The results from the two tests are compared at one point in time (cross-section), and statistical measures of diagnostic accuracy of the index test are calculated.

Example: In infants who undergo operative treatment to reduce a dislocated hip, we want to determine whether ultrasound is a good enough diagnostic tool in determining the position of the femoral head in the acetabulum, which indicates a successful operation. MRI can visualize the infant hip and can provide this information with a high degree of accuracy. However, MRI is expensive, warrants sedation of the patient, and is not easily available. In conducting the study, we include all infants who undergo the operation in a certain period. Each infant obtains, according to a standardized protocol, ultrasound and MRI tests in a consecutive sequence at the same time. When all patients have been included, we compare if the ultrasound results
Overview of Research Designs Applied to Radiology

Type of cases (population): Infants who undergo surgery to reduce a dislocated hip.
Type of controls (population): Not applicable.
Type of design: Cross-sectional (diagnostic test).

Cohort Design

In a cohort study the sample is studied over a certain period of time. Unlike in cross-sectional studies, time is an important aspect of cohort and case-control studies which aim at investigating the association between causes and outcomes (disease) or disease/procedure prognosis. In a cohort study, participants, none of whom has experienced the outcome of interest, are assembled based on certain inclusion criteria. They are followed for a predefined period of time after which the presence or absence of the outcome is ascertained (Fig. 1.5). This design is powerful in studying the association between certain variables that could be causally related to the outcome of interest. In diagnostic test research, cohort studies are useful when it takes time for a disease to manifest itself in

provided the same information as MRI about the location of the femoral head. That is, we are interested if ultrasound is as accurate as MRI in determining the position of the femoral head in the acetabulum after the operation.

The index test (ultrasound) and the reference standard (MRI) are applied at the same time and in all participants. The results of the index test are compared against those of the reference standard. In ideal circumstances, the results of the index test are consistent and almost identical to the results of the reference standard.

Research question: “Is ultrasound (index test) as accurate as MRI (reference standard) to determine the position of the femoral head in relation to the corresponding acetabulum in infants who undergo surgery to reduce a dislocated hip?”
Type of intervention: Not applicable (we are not comparing the efficacy of the surgical procedure).
Type of outcome: Position of the femoral head in relation to the corresponding acetabulum, best seen with MRI (reference standard).
a way that is detectable by diagnostic modalities or because of clinical symptoms. In cohort studies, the participants first receive the index test and, after a certain period of time, they receive the reference standard test. Follow-up time after the index test (latency period) is an important element in this design. Without it, the reference standard test cannot detect the disease of interest accurately.7

Cohort studies can be conducted prospectively (participants are assembled in the present and followed into the future) or retrospectively (the sample is identified from past medical records and followed forward from that point in time into the present). The latter are often performed using large health services databases such as electronic hospital admission records or insurance records.

**Example of a Prospective Cohort Design**

Investigators desire to determine the validity of ultrasound in the assessment of neonatal hip dysplasia. Clinicians consider radiography at the age of 6 months as the “reference standard” for this condition. The investigators set up a study in which all neonates born within a given period of time were included. Hip ultrasound, the index test, was applied to all participants at the age of 8 weeks. All participants were followed until the age of 6 months when they underwent a radiograph of the pelvis for assessment of hip dysplasia. The investigators compared the results of the ultrasound applied at the age of 8 weeks with the results of the radiography performed at the age of 6 months. If ultrasound is a valid test, its results will concur with the results obtained from radiography.

The index test was performed in all participants. It was known that it will take 6 months for the outcome of interest to occur in a way that is detectable by means of radiography (latency period). The reference standard was performed in all participants who had the index test done and it was performed after the index test.

**Research question:** “Is ultrasound (index test) performed at the age of 8 weeks in neonates able to predict the diagnosis of hip dysplasia established by MRI (reference standard) at the age of 6 months?”
Type of intervention: Not applicable. There is no intervention (clinical or surgical procedure) in the study. Ultrasound is the diagnostic test and MRI is the comparator.

Type of outcome: Presence of hip dysplasia at the age of 6 months.

Patients: Neonates with clinical suspicion of hip dysplasia who underwent an ultrasound scan at the age of 8 weeks.

Controls: Not applicable.

Type of design: Prospective cohort.

Example of a Retrospective Cohort Design

In pregnant women expecting babies in breech position in their third trimester of gestation, one academic center has been routinely performing intrauterine MRI examinations for the last 10 years. The investigators of this center set up a study in which they evaluated third trimester-pregnant women expecting babies in breech position. The purpose of this test is to assess details of the position of the lower extremities of the fetus that could be predictive of future developmental dysplasia of hips (DDH). Recent studies have raised the possibility that intrauterine MRI examinations performed in the third trimester of gestation can cause deafness in children at school age (hypothetical assumption). The investigators then decided to determine whether the fetal exposure to intrauterine MRI could have caused deafness in the children with clinical likelihood of hip dysplasia due to breech position who had an intrauterine MRI performed several years ago. Another group of children from a different center who were also in breech position but had an ultrasound instead of an MRI done in their third trimester of gestation were the control group. The investigators knew that there was a "point of exposure" (i.e., an intrauterine MRI examination), accrued a cohort of patients at the present time, and conducted auditive examinations on the patients who agreed to participate in the study, and compared the auditive examination results of this population with those of children from a different center who had instead a third-trimester intrauterine ultrasound examination. The major confusion here would be to consider this as a case-control design instead of a retrospective cohort design since in this example a control group was also available for the cohort design. The distinction between these two designs is that whereas in the retrospective cohort design (this example) the point of exposure is known, in a case-control design it is unknown.

Research question: "Is intrauterine MRI (index test) performed during the third trimester of gestation in fetuses in breech position (screening for hip dysplasia) able to cause auditive impairment in this population by the time they reach school age?"

Type of intervention: Not applicable (there is no intervention in this study).

Type of exposure: Intrauterine MRI performed during the third trimester of gestation in fetuses in breech position.

Type of outcome: Auditive impairment at school age.

Type of cases (population): Children of school age who underwent an intrauterine MRI in their third trimester of gestation.

Type of controls (population): Children of school age who underwent an intrauterine ultrasound in their third trimester of gestation.

Type of design: Retrospective cohort.

Because many more patients need to be enrolled than experience the event of interest in prospective cohort studies, these studies are typically expensive, in particular, for rare diseases (it will take a long time to sample the required number of participants) or if the participants need to be followed for a long time until the outcome occurs. Cohort studies have the important advantage that the exposure can be elicited (e.g., the result of a diagnostic test under investigation) without the bias that may occur if the outcome is already known. A bias is a systematic error, or deviation from the truth, in results or inferences.
Case-Control Design

In case-control studies participants who have developed the outcome of interest (the cases) are identified and their past exposure to suspected causal factors is compared with that of participants, who do not have the outcome of interest (the controls). Thus, a case-control study identifies the cases by their outcome status, such as the presence of a disease, but the point of exposure is unknown. This design looks backward to determine risk factors for a disease or for the effectiveness of a procedure. A key issue is that both cases and controls are sampled from the same population (Fig. 1.6). The case-control design can be employed for diagnostic accuracy studies. In this case, the reference standard test has already been performed and has identified the diseased participants before the index test is performed. Usually, all cases will undergo the index test but only a random sample of the controls. This approach will save costs. Such designs can be beneficial when the prevalence of a disease is low, or when the index test is costly or has potential side effects. For diseases with long latency periods (i.e., the time before it is detectable with the reference standard), the case-control design will save time and costs because it starts when the disease has already occurred. Reversing the order in which the index test and the reference standard are applied will not change estimates of diagnostic accuracy as long as all participants undergo both tests.

Example: Investigators plan to determine the diagnostic accuracy of ultrasound findings in early infancy in the diagnosis of hip dysplasia. Radiography at 2 years is a well-accepted reference standard for the diagnosis of this condition. A general ultrasound screening program is in place in a given region and all neonates of that region receive a hip ultrasound at the age of 6 weeks. The case-control study begins not in the neonatal age, but when these children are 2 years old. A radiograph of the pelvis is taken in all participants (reference standard) to establish the

Fig. 1.6 Design of a case-control study. Cases and controls are identified based on the reference standard test. The research looks back in time to determine the results of the index test in cases and controls. The association between the index and reference standard tests is then established.
cases at the age of 2 years—that is, those with hip dysplasia on radiographs. Investigators note that hip dysplasia is present in 17 cases and absent in 1900 controls. They evaluate the ultrasound images that had been taken at the age of 6 weeks in all 17 cases and in a randomly chosen subsample of 200 controls. They compare how ultrasound images taken at 6 weeks differed between cases and controls to determine whether specific findings on ultrasound were only seen in cases but not in controls.

The case-control design made use of data that was already available (it had been collected for routine clinical care), and the outcome of interests had already occurred (the study bypassed the 2-year period between index test and reference standard results). With random sampling, only a portion of the non-cases needed to be evaluated (time and cost savings) with a valid estimate of specificity at the expense of a loss in precision (wider confidence intervals). Of course, such a design relies on good quality and completeness of the ultrasound data.

### Experimental Clinical Research

#### Randomized Controlled Trial Design

A randomized clinical trial is a prospective study that aims at investigating the efficacy of at least two different interventions. The intervention could be a medication, an operation, a diagnostic test, or a diagnostic strategy. Patients are identified for eligibility by use of predefined inclusion criteria. Those consenting to participate in the trial are randomly allocated to either of the interventions—it is no longer the researcher who decides which participant ends up with which intervention. The principle of random allocation ensures that the patients allocated to each of the possible intervention groups will be roughly equivalent and therefore any effect observed between the groups can be linked to the effect of the intervention under investigation rather than to a characteristic of the individuals in the group. In other words, random allocation accounts for all differences in patient characteristics, including those who are unknown to the researcher.

In studies evaluating diagnostic tests, the randomized trial design is only useful when the test properties (reliability, validity) are well established; that is, only a test that has demonstrated a sufficient degree of accuracy will be tested in the trial. Typically, a trial in this context is exploring whether having the test will lead to better health outcomes (e.g., survival) compared to not having the test at all, or to having a different test. Clearly, no diagnostic test itself will benefit a patient—it is the subsequent action that follows the test result that will. Often the action is the treatment given to the patient if the result of the test is positive. Hence, most such trials actually study different diagnostic strategies than just diagnostic tests, which makes it essential that the actions promoted by the test result are outlined exactly before the trial begins and that all patients receive these actions based on their test results. Randomized trials are the most powerful way to deal with confounding by indication; that is, bias encountered when the allocation of a diagnostic test is unbalanced and based on the risk of future health outcomes (see Chapter 6, Randomized Controlled Trials). Randomized trials of diagnostic tests are commonly used in evaluating a diagnostic test as a potential screening tool.

**Example:** Investigators aim to determine the diagnostic value of ultrasound in the diagnosis of hip dysplasia. Ultrasound is the established diagnostic test for assessing infant hip dysplasia with a sufficient degree of accuracy in all patients, including those with very mild and very severe disease. Those undergoing an ultrasound test will have, on average, an earlier diagnosis made and this will enable clinicians to commence treatment earlier than usual. To determine if ultrasound screening of all newborns leads to better outcomes (e.g., fewer cases with advanced disease that requires operative treatment), newborns are randomized to (i) ultrasound screening or (ii) no ultrasound screening. Both groups will be treated as per a predefined protocol and the groups will be compared at 3 months of age by means of hip radiography. Ideally, in the group screened by ultrasound, the diagnosis is established earlier and treatment is initiated earlier, which should result in better outcomes.
The key aspects of this study are as follows: If the newborns are randomly allocated, then the groups are similar, on average, in all characteristics (gender, age, firstborns, mode of presentation, mode of delivery, race, etc.) that could possibly affect outcomes. This principle of random allocation will ensure that any difference in outcomes will be a result of the trial interventions—that is, the earlier-than-usual diagnosis obtained from the ultrasound test. The trial does not stop when the test results of the ultrasound are available; in order for the novel diagnostic modality (ultrasound for all newborns) to be useful in clinical care, it has to be connected with improved patient outcomes, which are to be established in this trial.

Research question: “Does the use of ultrasound as a screening imaging test for hip dysplasia in the neonatal period result in better morphology of the hips detected at 3 months of age by radiography?”

Experimental intervention: Diagnosis of hip dysplasia obtained from ultrasound in the neonatal period.

Standard intervention: Diagnosis of hip dysplasia obtained from radiography at 3 months of age.

Primary endpoint (outcome): Radiographic hip morphology at the age of 3 months.

Choosing the Appropriate Design

User’s guides to the medical literature abridge the various research designs into four major categories: therapy (randomized controlled trial design), diagnosis (prototype: cross-sectional design, but also cohort and case-control designs), harm (prototype: randomized controlled trial design, but also cohort and case-control designs), and prognosis (cohort design).

1. Early stages of the evaluation of a diagnostic test include studies on reliability and validity. Reliability studies are cross-sectional studies as there is no temporality involved. Studies of diagnostic accuracy can utilize cross-sectional designs as long as the reference standard test can be performed within the same time frame as the index test. These studies are considered inexpensive compared to studies that need to follow patients over time. In a cross-sectional design, several patients can theoretically complete all study maneuvers within a day.

2. If the outcome of interest, such as a disease, will take time to manifest itself in a way that it can be detected with a reference standard test, follow-up of the patients after they have undergone the index test is mandatory. The follow-up time (latency period) is usually long (months or years) and a cohort study will be required to detect developing disease subsequent to the index test.

3. Case-control studies are useful when the disease of interest is rare or when the latency period is long. In this instance, the participants are selected by their disease status, which has already been determined (with use of the reference standard) when the study is commenced. Using existing health record data, cases and controls are compared for their index test results in a retrospective fashion. The design is prone to bias, in particular as it relies on routinely collected health data.

4. In evaluating the impact of diagnostic tests on clinical outcomes, the key question is whether testing (or a novel test strategy) leads to improved health outcomes compared to no testing (or an established test strategy). Such questions are best studied in a randomized clinical trial to avoid various sources of bias.

Each research question can be addressed by means of multiple designs, from those with minimal safeguards against bias (case report or case series) to those with considerable safeguards (randomized controlled trials or systematic reviews/meta-analyses) (Table 1.2). A summary of characteristics and applications of research designs is shown in Table 1.3.
### Table 1.2  Research designs that can be used to address different research questions

<table>
<thead>
<tr>
<th>Element to be evaluated</th>
<th>Research question</th>
<th>Study design</th>
</tr>
</thead>
</table>
| Accuracy of test        | To what degree can the test differentiate between patients who have the condition of interest (target condition) and those who do not? | Cross-sectional study  
Delayed-type cross-sectional study  
Prospective cohort study |
| Impact of test          | Will the application of the diagnostic test lead to better outcomes compared to the nonapplication of the test or an alternative test? | Randomized clinical trial  
Systematic review/meta-analysis |

### Table 1.3  Characteristics and applications of research designs

<table>
<thead>
<tr>
<th>Study design</th>
<th>Key characteristics</th>
<th>Temporality</th>
<th>Applications</th>
<th>Advantages</th>
</tr>
</thead>
</table>
| Cross-sectional            | Index test and reference standard test are carried out at the same time; there is no follow-up period | All study maneuvers are performed at one point in time | Best type of study for determining the prevalence of a disease in a population and for judging diagnostic procedures; used in studies to determine the reliability and concurrent criterion validity | • Quick  
• Economic |
| Cohort                     | Index test is performed before the reference standard test, with a latency period between the two; there is a follow-up period for the sample | Prospective or retrospective | Best type of study for understanding the causes of a disease and the risk factors—best observational study; same group is followed prospectively or retrospectively; used in studies of diseases that develop over time before they can be detected with sufficient accuracy by the reference standard | • Enables calculation of incidence, relative risk and attributable risk  
• Enables estimation of time between exposure and outcome  
• Subjects who present with the outcome can be matched to controls |
| Case-control               | Patients are selected based on the result of the reference standard test             | Retrospective     | Best type of study for understanding rare diseases; used in studies of diagnostic accuracy of diseases with a long latency period between exposure and outcome | • Small sample size required  
• Quick  
• Economic  
• May generate preliminary data for a later, more complete study |
| Randomized controlled trial (RCT) | Patients are randomly allocated to two or more groups of diagnostic strategies | Prospective       | Best type of study for assessing efficacy of a treatment or a procedure; used in studies to test the implication of diagnostic tests or diagnostic strategies on patient outcomes | • Strong statistical design: random assignment to groups—confounders are equally distributed  
• Unbiased  
• Blinding more likely  
• Strong causal relationship |
Examples of imaging-related research topics in the radiological field are:

- Screening: early detection of diseases
- Diagnosis and staging of diseases
- Value of imaging-guided therapy vs. conventional treatment
- Diagnostic performance of new imaging technique vs. conventional technique
- Evaluation of response to treatment: Imaging as a secondary outcome measure in clinical trials of therapeutic agents or interventions
- Imaging as a predictor of clinical outcomes of diseases

### Highlights of Key Points

- Within the spectrum of nonexperimental/quasi-experimental design, diagnostic test and observational designs (cross-sectional, cohort, and case-control studies) are included. Within the spectrum of experimental design, randomized controlled trials (RCTs) represent the primary study category.

- One should use the PICO (people, patients, or population; intervention; control or comparison; and outcome) and FINER (feasible, interesting, novel, ethical, and relevant) approaches to formulate appropriate research questions.

- Whereas retrospective cohort designs do have a known point of exposure, case-controls do not. In case-control designs, cases are selected based on the outcome (disease [case]) and compared with another group (no disease [control]). In cohort designs, cases are not selected based on having or not having the disease (outcome) at the beginning of the study but the outcome is expected to occur for some patients (cases who have been exposed to something) but not to others (controls) during the conduct of the study.

- User’s guides to the medical literature abridge research designs into four major categories: Therapy (randomized controlled trial design), diagnosis (prototype: cross-sectional design, but also cohort and case-control designs), harm (prototype: randomized controlled trial design, but also cohort and case-control designs), and prognosis (cohort design).

### References

2 Descriptive Statistics

Mateen Shaikh and Hadas Moshonov

Learning Objectives

This chapter exposes the reader to several basic ideas used in summarizing quantitative data. They can broadly be categorized numerically and graphically.

- Graphical summaries
- Bar graphs
- Histograms
- Boxplots
- Scatterplots
- Numerical summaries
- Single points
  - Mean
  - Median
  - Quantiles
- Measures of spread
  - Mean absolute deviance
  - Variance
  - Standard deviation
  - Interquartile range

Introduction

This chapter refers to basic descriptive statistics. Although these are considered introductory level topics, they are very important for two reasons. First, they form the building blocks for more advanced methods. Second, they suggest which more advanced methods are (or are not) suitable, as more advanced methods are less applicable to a variety of problems than simpler methods, which provide cursory insight on almost any type of data.

This chapter forgoes the rigorous statistical definitions behind most descriptive statistics, and instead focuses on why the concepts are important, and how to calculate, understand, and apply them. To this end, this chapter shows several examples and, where appropriate, throughout the concepts section, provide the accompanying R code.

This chapter weaves visual and quantitative methods for describing data. The choice between using each is often simple: for accuracy, quantitative methods of summarizing data are preferred, but to simply convey relationships, graphical approaches are typically more useful. Importantly, the goal of a graphical method is not to convey the exact numerical information, though they may be inferred; their purpose is to show relationships.

Before beginning with statistics, it is important for the reader to grasp some knowledge on different types of data. The reader should refer to the chapter on measurements (Chapter 4) concerning distinctions between discrete and continuous data, as well as the properties of interval ordinal and nominal data.

Bar Graphs and Histograms

The bar graph (also called a bar plot) is a staple of presenting discrete information and is the foundation for the histogram, discussed later in this section. It provides relative information identical to what is conveyed in a two-column table. Consider the number of examinations conducted by type of imaging modality in a hospital as shown in Table 2.1.

Information is read in the table or viewed with the bar graph produced with the R code above. The resulting histogram is shown in Fig. 2.1.

Both the table and the bar graph convey the same information. When there are many categories, it is easier to see which bars are
very high, very low, or near a particular value than it is to scan through a table and obtain that information. Bar graphs are suitable for any data that are discrete—limited only by the convenience of plotting a certain number of categories. A bar graph with a million distinct categories provides the same information as a million-rowed table, but neither would be particularly illuminating.

In a bar graph, the order of the bars is essentially irrelevant since bar graphs are useful for nominal data. When the data are ordinal, however, a natural ordering of the bars becomes useful. For instance, consider the number of MRIs on the days of a particular week as shown in Fig. 2.2.

The bars can be rearranged in any order, as long as bars are labeled the same. The same information is displayed but this would clearly impede communication by being unintuitive. Therefore, the ordering of bars improves communication. A similar type of graph to the bar graph, which exclusively uses ordered data by agglomerating discrete ordinal data or discretizing continuous data, is the histogram.

Consider the following set of body temperatures:

38.8, 36.9, 36.5, 39.2, 38.6, 38.6, 38.8, 39.5, 40, 38.6, 38.4, 39, 36.6, 35.9, 39.5, 38.8, 37, 38.7, 38.8, 38.8, 38.1, 39.1, 37.2, 37, 39.3, 37.1, 38.7, 39.1, 37.4, 38.4, 36.7, 38.3, 39.1, 39.1, 38.5, 39, 36.9, 38.1, 38.7, 38.7, 36.8, 39.4, 38.7, 38.8, 39.2, 37.2, 39, 38.5, 38.6, 38.2, 39, 39, 37, 37, 39.1, 39.8, 37.1, 36.5, 39.2, 37.4, 37.8, 36.8, 36.5, 38.7, 39.2.

It is difficult to get a sense of the data when simply listed this way, but the data can be discretized into categories. Listing the frequency of observations falling in each category
Descriptive Statistics

There is a default heuristic used, but another choice can also be selected. Details of the other choices are described in the help file for the hist command: ?hist. The width of each bin can be specified in R using the `breaks=` parameter. The argument may be the endpoints of bins, or the approximate number of bins desired. Usually, R’s default choice is sufficient but it can be specified. For instance, consider the code that produces the histograms in Fig. 2.4.

Histograms provide a good approximation to how data spread or congregate within a data set. In this example, there are many cases where temperature is close to 39 degrees. This impression is easily obtained using the histogram but much more difficult to glean looking at the raw numbers. In addition, the range of data is easily seen but this is more difficult to obtain from looking at the raw data.

Referring back to Fig. 2.3, the data have two “bumps” in the histogram called modes. This makes the histogram bimodal, sometimes an indication of subpopulations. In this case, there is a group of people with higher body temperatures around 39 degrees and another group with lower temperatures, closer to 37 degrees. This may seem like a simple

---

**Table 2.2** Discretized body temperature values

<table>
<thead>
<tr>
<th>Range</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.5–35.9</td>
<td>1</td>
</tr>
<tr>
<td>36.0–36.4</td>
<td>9</td>
</tr>
<tr>
<td>37.0–37.4</td>
<td>10</td>
</tr>
<tr>
<td>37.0–37.4</td>
<td>1</td>
</tr>
<tr>
<td>38.0–38.4</td>
<td>6</td>
</tr>
<tr>
<td>38.5–38.9</td>
<td>18</td>
</tr>
<tr>
<td>39.0–39.4</td>
<td>16</td>
</tr>
<tr>
<td>39.5–39.9</td>
<td>3</td>
</tr>
<tr>
<td>40.0–44.4</td>
<td>3</td>
</tr>
</tbody>
</table>
10
0

5

Frequency

15

20

Research Methods in Radiology

36

37

38

39

40

Body temperatures (Celsius degrees)
Fig. 2.3 Histogram of body temperatures.
> temperatures=c(38.8, 36.9, 36.5, 39.2, 38.6, 38.6, 38.8,
39.5, 40, 38.6, 38.4, 39, 36.6, 35.9, 39.5, 38.8,
37, 38.7, 38.8, 38.8, 38.1, 39.1, 37.2, 37, 39.3,
37.1, 38.7, 39.1, 37.4, 38.4, 36.7, 38.3, 39.1,
39.1, 38.5, 39, 36.9, 38.1, 38.7, 38.7, 36.8, 39.4,
38.7, 38.8, 39.2, 37.2, 39, 38.5, 38.6, 38.2, 39,
39, 37, 37, 39.1, 39.8, 37.1, 36.5, 39.2, 37.4,
37.8, 36.8, 36.5, 38.7, 39.2)

4
2

3

Frequency

20
0

0

1

10

Frequency

30

5

40

6

> hist(temperatures,main=“”, xlab=“Body temperatures in degrees Celsius”)

34

35

36

37

38

39

40

Body temperatures (Celsius degrees)

36

37

38

39

40

Body temperatures (Celsius degrees)

Fig. 2.4 Two histograms of the same data with different bin widths.
> hist(temperatures,breaks=2,xlab=“Body temperatures in degrees Celsius”)

16

> hist(temperatures,breaks=50,xlab=“Body temperatures in degrees Celsius”)

Doria_CH02.indd 16

10/3/17 10:14 AM


Descriptive Statistics

Measures of Centrality

When presented with data, it is convenient to try to reduce them to the equivalent of a single number since a single number is easily understood. Further, when there are different collections of data, reducing each collection to a single number is useful for comparison purposes. For example, if the femur length in adults is generally larger than the femur length in young children, it is convenient to see this as a comparison of two numbers. There are several ways to represent a data set with a single number, and the most common are considered in this chapter.

Mean

The mean is undoubtedly the most common method of representing an entire data set by one number. It is also called the arithmetic mean, or average, and is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Its interpretation is also very simple. If every observation contributed equally to the data, what would the value of each observation be so that we obtained the same sum? Consider the number of imaging examinations produced by the radiology departments of

remark but it is a very important one: much of the statistics that follow in this book assume the data are or require that the data to be unimodal. In an example of unimodal data, consider the histogram of average diameter of breast tumors in Fig. 2.5.

This histogram is unimodal but has a very distinct shape. One side extends farther from the peak than the other. This indicates that although most tumors measure between 10 and 20 units, tumors can be rather large, up to 30 units, but most lie in the range of approximately 10 to 20 units. This is an example of another trait of histograms, called skewness. If unimodal data are symmetric, both the left and right ends of the histogram, called the tails, look approximately like mirror images of each other. If one tail extends farther, the histogram is said to be skewed in that direction. It is right-skewed (positively skewed) if the right (more positive) tail extends farther from the middle than the left tail. Conversely, the histogram is left-skewed (negatively skewed) if the left (more negative) tail seems to extend farther away from the middle than the right tail. As a final note, rigorous mathematical definitions exist for skewness, and it does indeed quantify this concept. For our purposes, however, it will be sufficient to be less detailed since most applications of skewness need only be as detailed as visually inspecting a graph.

■ Measures of Centrality

Fig. 2.5 Average breast tumor diameters illustrating a right-skewed histogram.
three different hospitals in the same city on a
given week: 110, 150, 250. The average num-
ber of imaging examinations in the city was
170. That is, if each department produced
exactly 170 imaging examinations, the city
would have still produced the same number of
examinations. This was calculated as follows:

\[
\text{average number of imaging examinations} = \frac{110 + 150 + 250}{3} = \frac{510}{3} = 170.
\]

In R, the following command finds the mean.

```r
> images <- c(110, 150, 250)
> mean(images)
## [1] 170
```
or written in one line

```r
> mean(c(110, 150, 250))
## [1] 170
```

One interesting note about the mean is that
it will always lie in the range of the data: it will
never be smaller than the smallest value or larger
than the largest value—it must be somewhere in
the middle, which fits with the notion of being a
measure of centrality, or the center of the data.

Consider the number of imaging examina-
tions by hospital again, but, instead, one hospi-
tal produces many more images—namely, the
hospital that previously produced 250 images
now produces 2,500 imaging examinations.

\[
\text{average number of imaging examinations} = \frac{110 + 150 + 2500}{3} = \frac{510}{3} = 1010.
\]

```r
> mean(c(110, 150, 2500))
## [1] 920
```

Although only one hospital produced more
than 200 imaging examinations, the average
was over 1,000. This demonstrates that re-
placing any data point with a larger point will
increase the average, and the converse is true:
replacing any data point with a point that is
smaller will decrease the average. The mean
is considerably influenced by extreme values.
This influence is one reason that the median,
discussed later, is another commonly used
measure of centrality.

A final note about means is that they are de-
\[\text{defined for a variety of types of data, although their}\]
\[\text{interpretation is not always so clear. Certainly}\]
\[\text{for continuous and even counts, the average}\]
\[\text{makes some sense but the average of ordinal}\]
\[\text{data from, say, a pain scale, may be somewhat}\]
\[\text{dubious. For instance consider asking patients}\]
\[\text{with suspected appendicitis to rate their pain}\]
\[\text{on a 5-point scale (0 = no pain, 5 = most intense}\]
\[\text{imaginable pain). Suppose the “average pain”}\]
\[\text{felt by individuals with appendicitis is known}\]
\[\text{to be 3.5. One could understandably be unsure}\]
\[\text{about how to interpret that value to compare}\]
\[\text{against individual responses. One particular}\]
\[\text{issue is that general arithmetic is often nonsen-
]\]
sical for ordinal data. For statistical analysis,
\[\text{the coarse level of the values is often an indica-
}\]
tion to use values such as medians (discussed
\[\text{later) rather than means, but other sources}\]
\[\text{have upheld that this is not such an issue for}\]
\[\text{the purpose of analysis, although interpreta-
}\]
tion can still be problematic. A more thorough
treatment on techniques used for such scales
\[\text{can be found in Streiner et al.}\]

\[\text{Weighted Means}\]

One generalization of the arithmetic mean is
the weighted arithmetic mean, which is very
simple to motivate. Consider the number of
imaging examinations produced over a larger
area, including clinics, in Table 2.3.

\[\text{Table 2.3 Frequencies of imaging examinations}\]

<table>
<thead>
<tr>
<th>Number of imaging examinations</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>31</td>
</tr>
<tr>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>200</td>
<td>16</td>
</tr>
<tr>
<td>500</td>
<td>7</td>
</tr>
<tr>
<td>1,000</td>
<td>4</td>
</tr>
</tbody>
</table>
The mean number of imaging examinations is

\[
\frac{31 \times 50 + 15 \times 100 + 16 \times 200 + 7 \times 500 + 4 \times 1000}{73} = \frac{13750}{73} = 188.35616\ldots
\]

Notice how it almost looks like the mean of just the numbers 50, 100, 200, 500, and 1,000, but instead of just adding these numbers, each number is multiplied by the frequency before adding. The frequencies in this case are used because they represent the same calculation. The above expression can be further manipulated to

\[
\frac{31}{73} \times 50 + \frac{15}{73} \times 100 + \frac{16}{73} \times 200 + \frac{7}{73} \times 500 + \frac{4}{73} \times 1000
\]

Each of the numbers in front of the amounts are called weights, and because all of the weights add to 1 \((\frac{31}{73} + \frac{15}{73} + \frac{16}{73} + \frac{7}{73} + \frac{4}{73} = 1)\), this is called a weighted mean. Note that the plain arithmetic mean is a special case of the weighted mean—namely, one where the weights are always \(\frac{1}{n}\), where \(n\) is the number of data points.

One of the most important instances when a weighted mean arises is in calculating an expected value. Consider the example counting the number of images from three departments (110, 150, 250 images). The middle point when data are ordered is 150. This differs from the mean, but like the mean it lies within the range of the data, and this will always be the case. To an extent, this is more so the case, because if the median of the data, 250, is replaced with 2,500, the median remains unchanged, the middle number remains 150. In this respect, the median is said to be more robust to extreme values, provided the order remains the same. This can be calculated in R as follows.

\[
\text{> median(images)}
\]

MEDIAN

The median is another way to obtain an impression of the “middle” of the data in another sense. Consider the example counting the number of images from three departments (110, 150, 250 images). The middle point when data are ordered is 150. This differs from the mean, but like the mean it lies within the range of the data, and this will always be the case. To an extent, this is more so the case, because if the median of the data, 250, is replaced with 2,500, the median remains unchanged, the middle number remains 150. In this respect, the median is said to be more robust to extreme values, provided the order remains the same. This can be calculated in R as follows.

\[
\text{> median(images)}
\]

Median

The median is another way to obtain an impression of the “middle” of the data in another sense. Consider the example counting the number of images from three departments (110, 150, 250 images). The middle point when data are ordered is 150. This differs from the mean, but like the mean it lies within the range of the data, and this will always be the case. To an extent, this is more so the case, because if the median of the data, 250, is replaced with 2,500, the median remains unchanged, the middle number remains 150. In this respect, the median is said to be more robust to extreme values, provided the order remains the same. This can be calculated in R as follows.
or written in one line

> median(c(110,150,2500))
## [1] 150

The median is easy to find when there are an odd number of data points: with three points we take the second highest value since that would be the middle. If we have 11 points, the sixth highest value is taken. In general, if we had \( n \) points, we would take the point corresponding to the \( \frac{n+1}{2} \) highest. For odd \( n \), \( n + 1 \) is an even number, so dividing by two gives a whole number. When \( n \) is even, \( n + 1 \) is odd, and dividing by two will give a number with the decimal .5 at the end. This requires that we adjust the technique.

Consider one more department in our example so that the numbers of imaging examinations produced are 110, 150, 250, 300. There are now two values equally close to the middle: the second highest and third highest. Because the median gives a sense of the middle, a number in the middle of these two numbers is chosen. Usually, we take the average of the two, or the midpoint of the two numbers. Note that in this case, we have ended up with a middle number that is not even in the data set:

> images <- c(110,150,250,300)
> median(images)
## [1] 200

or written in one line

> median(c(110,150,250,300))
## [1] 200

The influence of extreme points can be well seen with a histogram. Consider the histograms in Fig. 2.6 with more values like those in the previous example. In the first histogram, data are approximately symmetric and the mean and median are approximately the same value. In the second histogram, the largest 40% of data points have been made slightly larger. The dotted line indicates the mean whereas the solid line indicates the median. The median remains completely unchanged even though nearly half of the data points have been changed, but the mean does change.

![Histograms showing impact of extreme values on mean and median.](image_url)

**Fig. 2.6** Impact of extreme values on the mean and median.
Descriptive Statistics

Illustrated graphically. Consider the following example of the number of imaging examinations from a small region:

100, 100, 110, 125, 130, 150, 150, 150, 175, 175, 175, 175, 200, 200, 250

The most frequent number is 175, and so the mode is 175. Like the median and mean, a mode cannot be smaller than the smallest value or greater than the greatest value, so it is a measure of centrality in that sense. This is verified in a histogram of the data shown in Fig. 2.7.

Mode

The mode is probably the least used of the three measures of centrality discussed in this chapter. It is not a complicated concept: mode is the most frequent number to appear in discrete data. Unfortunately, this is not always a useful concept to be applied in the previous example: every number appeared the same number of times, which is very common in samples from continuous data. In these cases, the mode is still defined as the number seen most often (if the sample was large) but is not helpful for smaller sample sizes. As an approximation, the data can be discretized into coarse bins just like in histograms, and some value within the most frequent bin, such as the midpoint, is taken to be the mode. More rigorous approaches have also been developed to calculate modes in such cases. Using the histogram approach, the mode is illustrated graphically. Consider the following example of the number of imaging examinations from a small region:

100, 100, 110, 125, 130, 150, 150, 175, 175, 175, 175, 200, 200, 250

The most frequent number is 175, and so the mode is 175. Like the median and mean, a mode cannot be smaller than the smallest value or greater than the greatest value, so it is a measure of centrality in that sense. This is verified in a histogram of the data shown in Fig. 2.7.

Measures of Quantiles

The section about medians briefly discusses the median as representing the middle data point, if data were ordered. In fact, ordering data is a very important aspect of statistics that corresponds to analyzing quantiles. Imagine ordering all of the data along a line and choosing a data point somewhere between the beginning and the end. If the halfway point is between 100 and 150, for example, the data can be divided into two equal parts: those less than or equal to 125 and those greater than 125. If there were an even number of data points, the halfway point would be the average of the two middle-most numbers.

All that is required for finding a median is ordering the data to find out what the middle-most number is, or in the case of an even number of data points, taking the average of the two middle-most numbers.

Fig. 2.7 Number of imaging examinations in a small region with a mode of 175 images.
point is chosen, the selected datum is the median. The halfway point can be referred to as the number 0.5; therefore, the quantile for 0.5 is the median. Any number can be chosen, such as a point that is one quarter the way or three quarters of the way from the beginning to the end. These would correspond to the quantiles 0.25 and 0.75.

```r
> numImages <- c(100,130,175,200, 100,150,175,200,110,150,175,250, 125,150,175,250)
> median(numImages)
## [1] 162.5
> quantile(numImages,0.5)
## 50%
## 162.5
> quantile(numImages,c(0.25,0.75))
# calculates a collection of two quantiles
# at once with c(0.25,0.75)
## 25% 75%
## 128.75 181.25
```

Some proportions are commonly used. The median was designed to split the data in two halves. If the data are cut into thirds, the corresponding quantiles would be tertiles. If the data are cleaved into quarters, the corresponding quantiles are called quartiles. It is also common to consider the data being split into 100 sections, creating percentiles. The 89th percentile could also be calculated as 217.5. Note that in this case, like the case of finding a median from an even number of data points, the value of the quantile is estimated from one of many approaches beyond the scope of this book. Methods of estimating quantiles are available in most data analysis software.

**Boxplot**

A histogram conveys quite a bit of useful information, and its analysis often occurs when data are unimodal, or have a single mode. In this instance, most of the summary information in the sample is obtained by looking at the range, the median, and two matching quantiles, such as the first and third quartiles. When these five pieces of information are all someone wishes to use to obtain an impression of data, this can be easily represented in a boxplot, shown in Fig. 2.8.

The boxplot is also useful for skewed, unimodal data. The ends of the box correspond to the first and third quartiles, with the mark in the middle of the box indicating the second quartile, the median. One indication of skewness comes from the line in the middle, the median, not lying in the middle of the first and third quartiles. Examples of this are shown in Fig. 2.9a and Fig. 2.9b. There are lines extending from the box called whiskers. The length of whiskers follows less universal rules. Whiskers never extend beyond the range of the data and are meant to indicate where another large mass of the data lies. In R the default is to make the whiskers no more than 50% longer than the distance between Q1 and Q3, consistent with a common convention by the boxplot’s inventor.10 Any points beyond this range are explicitly drawn as being of particular interest, possibly outliers. Such points are few in number (relative to the size of the data) in unimodal data. It is possible that these points represent different populations or measurement errors, but given large enough samples, it is expected that some points will be far away from the middle of the data.

When data are not unimodal, impressions of data are extremely unreliable. Recall our example of the duration of time to conduct imaging examinations. The boxplot is shown in Fig. 2.10.

The boxplot, a summary of five numbers, is not capable of describing the nuances of multimodal data. Splitting the data by type of imaging modality and providing a boxplot for the X-rays and MRIs separately provides a better summary. These are typically done without histograms, as shown in Fig. 2.11.
Fig. 2.8 Comparison of boxplot and histogram.

Fig. 2.9 Histograms of (a) left-skewed data and (b) right-skewed data.
The quartile is easily obtained with the R function `summary`, which prints the minimum, maximum, first quartile, second quartile (median), and third quartile. This summary of five numbers is called the five-number summary, and when a boxplot chooses whiskers that extend to the range of the data, it contains the exact same information.

```r
> summary(duration$length)
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 18.22 32.67 55.31 56.89 74.93 135.70

> summary(duration$length[duration$type=="X-Ray")
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 34.48 66.95 75.06 78.02 89.32 135.70

> summary(duration$length[duration$type=="MRI")
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 18.22 25.24 32.67 35.77 43.16 74.74
```

Fig. 2.10 Boxplots and histogram of the number of examinations produced altogether and by type of imaging modality.

Fig. 2.11 Boxplots of durations of examinations by type of imaging modality.

```r
>boxplot(duration$length~duration$type)
```
Measures of Spread

Although it may seem natural to use the range of the data to represent a measure of spread, this is often not useful as it is very sensitive to extreme observations, so other approaches are used. The spread of data is almost always measured using two related quantities: the variance or its square root, the standard deviation. These two quantities may seem less intuitive than means but are used for many analytical reasons. This is motivated by another quantity: the absolute mean deviation.

Consider first the average of a set of numbers. Rather than considering where the middle of the data is, consider how far away each of these numbers is from the middle. Consider an unrealistic but easy to follow example: the numbers 60, 70, 80, 90, 100, where the sample mean is 80. Now ask, on average, how far away each number is from the mean of 80? For each number, we calculate this difference, for each \( x_i \),

\[
x_i - \bar{x},
\]

which is called a deviation, because it indicates how much an observation strays from the mean. Therefore, what is the average value of the deviation? This is illustrated in Table 2.5.

```
> dat <- c(60,70,80,90,100)
> mean(dat)
## [1] 80
```

<table>
<thead>
<tr>
<th>Observation</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>-20</td>
</tr>
<tr>
<td>70</td>
<td>-10</td>
</tr>
<tr>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

This cannot be right, can it? On average, there is no distance between data and the mean. It is and it is not right. It is right in the sense that the mean was correctly calculated. Our interpretation is not right, because the distances were allowed to be negative. Instead of taking the average of the deviations, we should take the average of the absolute value of the deviations.

```
> deviations <- dat-mean(dat)
> deviations
## [1] -20 -10  0 10 20
> mean(deviations)
## [1] 0
```

That makes more sense. Observations are, on average, 12 units away from the mean. Before continuing, note that the average of (non-absolute value) deviations will always be zero, so making a deviation positive somehow is necessary before taking the average.

```
> absolutedeviations <- abs(deviations)
> absolutedeviations
## [1] 20 10  0 10 20
> mean(absolutedeviations)
## [1] 12
```

Often, it is useful to know how spread out the data are when comparing different data sets. Consider the following data set with the same mean: 20, 40, 80, 100, 120.

```
> dat2<-c(20,40,60,80,120)
> mean(dat2)
## [1] 64
```

```
> deviations2 <- dat-mean(dat2)
> absolutedeviations2 <- abs(deviations2)
> mean(absolutedeviations2)
## [1] 17.6
```
These data are more spread out, although they are centered around the same value.

**Variance and Standard Deviation**

Although it does measure the spread of observations around their mean, the mean of the absolute value of the deviations is infrequently used to quantify spread. Instead, the deviations are squared, and the mean of these squared deviations is calculated. The result is the variance.

```r
> squareddeviations <- deviations^2
> mean(squareddeviations)
## [1] 200
```

The units of variance are the square of the original units, so the square root of the variance is taken to produce a value on a scale comparable to the scale of the original data. The result is called the standard deviation. The formula below shows how to calculate it in R. There is a built-in function, `sd()`, that performs the same calculations.

```r
> sqrt(mean(squareddeviations))
## [1] 14.14214
```

The formal definition of the sample standard deviation is given by:

\[
s = \sqrt{\text{mean of squared deviations}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}
\]

and the variance, the square of the standard deviation, \( s^2 \), is therefore

\[
s^2 = \text{mean of squared deviations} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}
\]

A final note about variance and standard deviation is that in practice the denominator can be different from just the number of observations. The most common denominator to use is \( n - 1 \) rather than \( n \). The rationale for this is a desirable property called unbiasedness, meaning that the estimate \( s \) will on average be equal to the true value of the variance. The concepts of parameters and their estimators are covered in detail in Chapter 13 on statistical inference.

**Quantile-Based Spread**

As seen with boxplots, one can visualize how spread out the data are by looking at a pair of quantiles. The most common method of using quantiles to assess the spread of the data is through the interquartile range, or the distance from the first quartile to the third quartiles. This can be visually inferred (approximately) from the boxplot.

**Measures of Association Between Two Variables**

Thus far, the focus has been on a single variable. In the case where two discrete variables are recorded for each observation, the relationship between them can be effectively summarized in a table. Analysis of these tables is left for later chapters on diagnostic tests, measurements, and inference. Instead, the focus is now on cases where both measured variables are continuous.

Likely the most important characteristic of two variables is how they vary relative to one another. The classic example is height and weight. Often, tall people weigh more than short people. This does not mean that all tall people weigh more than short people, or that short people cannot be heavy, but there is some relationship between the two characteristics.

**Scatterplot**

The relationship between two variables can best be seen in a scatterplot. Consider the breast tumor data again. In addition to the tumor size, a measurement was made of fractal dimension, a variable related to how
Consider the multiple measurements from the breast cancer tumors: mean values of each tumor for:

- Diameter (3rd variable)
- Perimeter (5th variable)
- Area (6th variable)
- Symmetry (11th variable)
- Fractal dimension (12th variable)

The 10 possible individual pairwise scatterplots are arranged on the scatterplot matrix illustrated in Fig. 2.13.

The scatterplots on the top right contain the same data as the scatterplots on the bottom left, except that they have the x- and y-axes reversed. Some of the variables have strong associations with each other (diameter, perimeter, area) whereas the others have a relationship

Scatterplot Matrix

When more than two variables are being considered, it becomes difficult to display them well on a single two-dimensional plot. A common approach is to create all of the possible scatterplots of two variables at a time.

ragged (non-smooth) the tumor edges were. This is illustrated in Fig. 2.12.

First, ignore coloring and observe that, generally, the larger the tumor, the lower the fractal dimension. When considering the coloring of the scatterplot, notice that malignant tumors tend to be larger and have a greater fractal dimension at a given size, whereas benign tumors are smaller and have a smaller fractal dimension at a given size.

Fig. 2.12 Scatterplot of tumor fractal dimension versus tumor radius (tumor diameter = tumor radius × 2).

```r
> plot(wdbc$radius_mean, wdbc$fractal_dimension_mean,col=wdbc$diagnosis, xlab="Radius", ylab="Fractal Dimension")
> legend(legend = c("malignant","benign"), col = c("red","black"), x="topright",inset=.1,pch=1)
```
that is not as strong (symmetry, fractal dimension) but there is a difference in relationships between pairs of variables when cases are identified as malignant (red) or benign (black).

### Highlights of Key Points

This chapter focuses on various ways to describe data, both numerically and visually. Properties of data visually identifiable through a histogram:

- **Symmetry**
- **Skewness**
  - Left/negative skewness: tail extends farther to the left
  - Right/positive skewness: histogram tail extends farther to the right
- **Modality**
  - Unimodal: histogram is generally one-humped
  - Bimodal: histogram is two-humped
  - Multimodal: histogram has multiple humps

Numerical methods of describing data:

- **Measures of centrality**
  - Mean or weighted mean: sum of data divided by number of data points
  - Median: half the observations are smaller than the median and half are larger
  - Mode: most frequent number in a data set

- **Measures of quantiles**
  - Quartiles: data points that cut the data into quarters
  - Boxplots: visual representation of quartiles, median, and extremes
  - Percentiles: data points that cut the data in 100 segments

---

**Fig. 2.13** Scatterplot matrix of five tumor features.

```r
> pairs(wdbc[,c(3,5,6,11,12)],col=wdbc$diagnosis)
```

---

Doria_CH02.indd 28

10/3/17 10:14 AM
• Measures of spread
  • Variance: average squared distance of observations from the mean
  • Standard deviation: square root of variance
  • Interquartile range: middle 50% of data

References

Diagnostic Tests
George Tomlinson, Gerald Lebovic, Connie Marras, and Andrea S. Doria

Learning Objectives

- To apply methods for estimation and hypothesis testing for proportions to the binary diagnostic test characteristics of sensitivity, specificity, and positive and negative predictive values.
- To apply methods for estimation of ratios of proportions to the estimation of positive and negative likelihood ratios.
- To construct receiver operating characteristic (ROC) curves for a test that has more than two diagnostic levels and estimate the area under the curve (AUC).

This chapter introduces key concepts involved in a study of a diagnostic test: sensitivity, specificity, and positive and negative predictive values of binary diagnostic tests, likelihood ratios, and the ROC curve. As we introduce these concepts, we demonstrate statistical methods for estimation of these quantities and their confidence intervals, and for hypothesis testing.

Summary Measures of Diagnostic Accuracy

To streamline the presentation of materials here, we need to introduce some terminology. First, define D as a binary variable representing the true disease status, which takes the value 1 if someone has the disease and 0 if not. This requires the existence of a reference standard that correctly identifies patients as having the disease or not. A reference standard is the test (or group of tests) best representing the true disease state. A classic example of a reference standard is the result of a biopsy with histologic examination of a pulmonary nodule seen on a computed tomography (CT) scan of the chest, which classifies it as benign or malignant. Another example of a reference standard is sufficiently long clinical follow-up of patients with clinical suspicion of appendicitis to determine the eventual development of appendicitis or freedom from this condition. In many cases, however, a reference standard is not available as it is either not ethical or not feasible to carry out the often invasive procedures needed to obtain certainty about the presence of disease. This leads to the situation where the best currently available diagnostic test (or combination of tests) is used as a de facto reference standard. For example, magnetic resonance imaging (MRI) or CT imaging may be used as a reference standard for a proposed diagnostic technique based on ultrasound.

Next, we define T as a binary variable that is the result of a diagnostic test and that takes the value 1 for a positive test and 0 for a negative test. In the assessment of medical images, the classification of an image as T = 1 versus T = 0 relies on the synthesis of information across the image, or even across several images or different imaging modalities. This information can often be thought of as taking on some value X, where high values are more suggestive of disease and lower values are more indicative of no disease. The variable X may be a measurable quantity, such as wash out time or a quantification of the degree of enhancement. But in many settings, X is not measured directly based on the image, but represents, for example, a degree of suspicion, a qualitative assessment of the level of enhancement, or a combination of characteristics associated with disease—opacity, size, and calcification, for example, in the case of MRI for detection of breast cancer. Whether X is quantified or X is a general measure of the amount of evidence for disease, if X
is used to produce a binary diagnosis of $T = 1$ or $T = 0$, it must be compared to a threshold, which we call $k$. This threshold is the minimum level of evidence that the reader of the images uses to make a diagnosis of disease. This contrasts with many medical diagnostic tests based on blood levels, where, for example, levels of serum creatinine above a certain level are compared to a strict numerical threshold (e.g., $> 1.5 \text{ mg/dL}[133 \text{ mmol/L}]$) to classify a patient having an MRI examination as being at high risk for gadolinium-induced nephropathy.\textsuperscript{4} Values of $X$ at or above the threshold result in a diagnosis of disease, $T = 1$; and values of $X$ below the threshold result in a patient being diagnosed as disease free, $T = 0$. That is, if $X \geq k$ then $T = 1$, and if $X < k$ then $T = 0$.

A person who has undergone both a reference standard assessment and the diagnostic test can be classified in one of four ways. Table 3.1 summarizes data from a binary diagnostic test study and can be used to estimate all the relevant quantities.

In this study design, each subject with disease is considered to have the same probability of being diagnosed positive. This probability of a positive diagnostic test, given that the subject has disease, is called the sensitivity (se) or true-positive fraction:

$$\text{sensitivity} = P(T = 1 \mid D = 1)$$

This is estimated from Table 3.1 as $\hat{se} = a/(a + c)$, the proportion of all diseased cases that have a positive test. The circumflex, or “hat,” over se indicates that it is not the actual sensitivity, but an estimate of sensitivity.

All subjects without disease on the reference standard are considered to have the same probability of obtaining a negative diagnostic test. This probability is called the specificity (sp), or true-negative fraction.

$$\text{specificity} = P(T = 0 \mid D = 0)$$

and is estimated from Table 3.1 as $\hat{sp} = d/(b + d)$.

The proportions of patients wrongly classified by the diagnostic test do not have special names, but are known simply as the false-positive fraction (for those healthy patients with a positive test) and the false-negative fraction (for the diseased patients with a negative test). These proportions are estimated from Table 3.1 as $FP = b/(b + d)$ and $FN = c/(a + c)$. Alternatively, they are found as the complement of specificity ($FP = 1 - \text{specificity}$) and sensitivity ($FN = 1 - \text{sensitivity}$).

The sensitivity of a test answers the question: Given that a subject has the disease ($D = 1$), what is the probability that the test will pick up the disease? Similarly, the specificity of a test answers the question: Given that a subject does not have the disease ($D = 0$), what is the probability that the test will be negative? These quantities examine probabilities by doing what is called “conditioning” on the true state of the disease. Whereas the true disease state is often (but not always) known in a study of a diagnostic test, in the clinical setting where the diagnostic test will be used, the true disease state will not be known. In fact, the reason for administering a diagnostic test is to obtain an improved estimate of the probability that a patient has the disease. In the clinical setting, the questions of interest are conditional on the result of the diagnostic test. Given that a diagnostic test is positive, the probability that a patient has disease is called the positive predictive value (PPV) and can be written as $P(D = 1 \mid T = 1)$. When the diagnostic test is negative, the probability that a patient is disease-free is called the negative predictive value (NPV) and can be written as $P(D = 0 \mid T = 0)$. In Table 3.1, these quantities are estimated simply as the
The proportion of patients with a positive test who have the disease, \( \hat{PPV} = \frac{a}{a + b} \), and the proportion of patients with a negative test who do not have disease, \( \hat{NPV} = \frac{d}{c + d} \).

Chapter 13 shows how to test a hypothesis and how to calculate a confidence interval—a range of plausible values—for a proportion, an approach that can be used for all of the summary proportions from Table 3.1. Keep in mind that all these equations refer to proportions, not percentages. Once all the calculations have been done on proportions, the resulting values may be presented as percentages.

### Example: Ultrasound for Assessment of Appendicitis

Schuh et al.\(^5\) examined the use of ultrasound scanning as a diagnostic tool to be used in the emergency department in children who have suspected appendicitis. The paper classifies the results of the ultrasounds into three broad categories: positive (95 children), negative (76 children), and equivocal (92 children). The positive and negative categories refer to cases where the sonographer was able to visualize the appendix and classified the test result as either positive or negative, respectively, for appendicitis. Those examinations that were unable to properly visualize the appendix or had nondiagnostic features were classified as equivocal ultrasounds. **Table 3.2**

<table>
<thead>
<tr>
<th>Ultrasound finding (T)</th>
<th>True state (D)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appendicitis</td>
<td>No appendicitis</td>
</tr>
<tr>
<td>Positive for appendicitis (T = 1)</td>
<td>86</td>
<td>9</td>
</tr>
<tr>
<td>Equivocal</td>
<td>13</td>
<td>79</td>
</tr>
<tr>
<td>No appendicitis</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>164</td>
</tr>
</tbody>
</table>

**Table 3.3** Grouping positive and equivocal ultrasound findings from Table 3.2 as positive

<table>
<thead>
<tr>
<th>Ultrasound finding (T)</th>
<th>True state (D)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appendicitis</td>
<td>No appendicitis</td>
</tr>
<tr>
<td>Positive or equivocal for appendicitis (T = 1)</td>
<td>99</td>
<td>88</td>
</tr>
<tr>
<td>No appendicitis (T = 0)</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>164</td>
</tr>
</tbody>
</table>

summarizes the results for the positive, equivocal, and negative groups.

If we try to apply the definitions developed in Table 3.1, we are quickly faced with a decision: what to do with the equivocal cases? We need a binary classification of \( T = 1 \) or \( T = 0 \) for each case. In the Schuh et al. paper,\(^5\) equivocal scans were classified two ways, as either false-positive (the 79 scans where the child did not have appendicitis) or false-negative (the 13 scans where the child did have appendicitis), but here, we will take a different approach. If equivocal patients are going to be treated as if the ultrasound finding was negative and the patient was sent home, then they should be classified with the \( T = 0 \) row of Table 3.2. If the patients are going to form a “possible or probable appendicitis” cohort that was sent for further treatment or follow-up, then they are better classified with the \( T = 1 \) patients. The latter situation is more appropriate here, so we collapse Table 3.2 to form the 2 × 2 table shown in Table 3.3.

We can estimate specificity as \( \hat{Sp} = \frac{76}{88 + 76} = 0.463 \). Sensitivity is estimated as \( 99/99 = 1.0 \).

Below, we use the `prop.test` function in R to compute 95% confidence intervals for sensitivity and specificity.

```r
> prop.test(99, 99, correct=F)
[output snipped]
95 percent confidence interval:
0.963 1.000
```
Sensitivity values as low as 0.963 lie in the confidence interval, even though we observed perfect sensitivity. Notice that the confidence interval is not symmetric.

> prop.test(76, 164, p=0.6, correct=F)

```
data: 76 out of 164, null probability 0.6
X-squared = 12.7, df = 1, p-value = 0.0003564
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
0.389 0.540
sample estimates:
p
0.463
```

The estimate and confidence interval are the same as in the analysis above because we are using the same observed data. But we are now testing a different hypothesis, so get a different p value; this result suggests that such low observed specificity would be unusual (p = 0.0003564) if specificity were as high as 60%, so we reject that hypothesis.

Estimation and construction of confidence intervals for the remaining proportions for the 2 × 2 table (FP, FN, PPV, NPV) can all use prop.test in R, no matter whether there are zero counts or not.

The occurrence of a zero count, as we saw for the number of false-positives in Table 3.3, is common enough that it deserves the attention we gave it above. However, it does somewhat complicate the presentation of some additional important concepts in the analysis of a binary diagnostic test, so we continue this section with the data in Table 3.4, formed by grouping the equivocals and negatives in Table 3.2 into an “ultrasound negative” group. This grouping is appropriate if those with a positive ultrasound undergo some invasive procedure that should not be used on equivocal cases.

Sensitivity and specificity in Table 3.4 are estimated to be 86.9% and 94.5%. In this sample, 86 of the 95 patients with a positive ultrasound are cases of appendicitis, so the PPV...
is 90.5%. Among those with a negative ultrasound, 155/168 do not have appendicitis, so the NPV is 92.3%. If a positive ultrasound was used to decide on further treatment, 90.5% of the treated patients in this sample would have appendicitis and 92.3% of untreated patients would be free of appendicitis. These numbers suggest that ultrasound can be a useful diagnostic tool to direct potential cases of appendicitis to the appropriate clinical management path. Can the readers of the Schuh et al. study\(^5\) apply these measures of diagnostic accuracy to their own emergency rooms (ERs)? This study took place in the ER of a pediatric hospital downtown in a major metropolitan area; it is possible that the prevalence of appendicitis there is not typical of other settings where ultrasound might be used to diagnose appendicitis. Would ultrasound be as useful in a setting with a lower prevalence? Let’s imagine a study where the number of children without appendicitis is 10 times as high—they have some other cause for their symptoms. These hypothetical numbers are presented in Table 3.5.

Clearly in Table 3.5, sensitivity is still 86.9% and it is easily verified that specificity is still 94.5%; the numerator and denominator of the estimated specificity are both 10 times as large as in Table 3.4. But now, only 86 of the 176 patients who are positive on ultrasound are cases of appendicitis, so PPV = 48.9%. Someone who applied the PPV from this study to a low-prevalence setting would be sending more than half of the children with a positive ultrasound for unnecessary treatment. By contrast, 1550 of 1563 patients with a negative ultrasound are free of appendicitis, so NPV = 99.2%, higher still than the value in Schuh et al.\(^5\)

While it is generally felt that the sensitivity and specificity of a diagnostic test are properties of the test (which includes contributions from the ultrasound device itself, the sonographer, and the reader of the diagnostic test in the case of ultrasound), the positive and negative predictive values are also dependent on the prevalence of disease. How can the results from a study carried out in a setting with one value for prevalence be applied to a setting with a different value for prevalence? This situation exists in primary care: the prevalence of most diseases is low, but diagnostic tests that will be applied there may have been developed and evaluated in a setting with a high disease prevalence.

The PPV in a new setting with prevalence \(P\) can be related to sensitivity \((se)\) and specificity \((sp)\) through the following formula:

\[
PPV = \frac{se \times P}{se \times P + (1 - sp) \times (1 - P)}
\]

The numerator represents the probability of a diseased patient testing positive: it is the proportion of patients with disease times the proportion of those who test positive. The numerator is the probability of a positive test, irrespective of true disease state: the probability of a diseased patient testing positive \((se \times P)\)
The previous section shows how to use the two summary statistics of sensitivity and specificity from one setting to calculate PPV and NPV in a setting with a different prevalence. It is also possible to summarize the value of a diagnostic test using two different summary statistics, the positive and negative likelihood ratios. The likelihood ratio of a positive test is defined as:

\[ LR^+ = \frac{P(T = 1|D = 1)}{P(T = 1|D = 0)} = \frac{\text{sensitivity}}{1 - \text{specificity}} \]

This ratio measures how much more likely a positive test is in a diseased patient than it is in a healthy patient. As we show here, it is also the relative increase from the pretest odds of disease in a patient who has a positive result on the diagnostic test.

In a similar fashion, we can define the likelihood ratio for a negative test as the relative probabilities of a negative test in diseased and nondiseased patients. This ratio will generally be less than one and measures the relative decrease from the pretest odds of disease in a patient who has a negative result on the diagnostic test.

\[ LR^- = \frac{P(T = 0|D = 1)}{P(T = 0|D = 0)} = \frac{1 - \text{sensitivity}}{\text{specificity}} \]

This ratio measures how much more likely a positive test is in a diseased patient than it is in a healthy patient. As we show here, it is also the relative increase from the pretest odds of disease in a patient who has a positive result on the diagnostic test.

Knowledge of the sensitivity and specificity of a test will help clinicians determine whether the test is most useful in ruling in or ruling out a disease. For example, a test that is very sensitive will rarely miss people with the disease and a negative result obtained from a test with high sensitivity will therefore be useful in ruling out disease (although a positive result will not necessarily rule in the disease). Conversely, a highly specific test will rarely misclassify people without the disease as diseased and a positive result will be useful in ruling in a disease (although a negative result in this situation does not necessarily rule out the disease).
the larger the increase in the probability that the person does not have the disease with a negative test.

In Table 3.3, the positive likelihood ratio is $LR^+ = 0.869/(1 - 0.945) = 15.8$ and the negative likelihood ratio is $LR^- = 0.139$. The odds of disease increase almost 16-fold after a positive test and decrease by a factor of around 7 after a negative test. As with sensitivity and specificity, we do not learn anything about the absolute odds of disease from the $LR^+$ and $LR^-$, only how much the odds change.

To better understand the use of likelihood ratios, it is necessary to be comfortable working with odds. The odds are the ratio of a probability for an event occurring to the probability against the event occurring. If $P$ is the probability that a patient has appendicitis, then the odds of appendicitis are $Odds = P/(1 - P)$. The equation that converts an odds into a probability is $P = Odds/(1 + Odds)$. Odds can be as low as zero but have no upper bound. For example, if the probability of an event is 0.10 (10%) then the odds are $0.10/(1 - 0.10) = 0.10/0.90 = 1/9 = 0.111$. If the probability of an event is 99% then the odds of the event are $0.99/0.01 = 99$, sometimes expressed as 99 to 1. It is important to keep in mind that doubling the odds does not mean doubling the probability. In the case where the odds are 99:1, doubling the odds means they go from 99 to 198 and the resultant probability is $198/(198 + 1) = 0.995$; we doubled the odds but increased the probability by only one half of one percent.

Before a test is performed, one can estimate a pretest probability, $P$, of the disease using all available information on the patient (e.g., age, gender, prevalence of the disease, previous diagnostic tests, etc.). The pretest odds, $Odds_{prior}$, can be found as $P/(1 - P)$.

The odds of disease after a positive test, $Odds_{post}^+$ can be calculated as:

$$Odds_{post}^+ = LR^+ \times Odds_{prior}$$

Likewise, the odds of disease after a negative test are equal to the product of the pretest odds of disease and the likelihood ratio negative

$$Odds_{post}^- = LR^- \times Odds_{prior}$$

For clinical use, where probability is a more familiar expression of risk than odds, these are usually converted back to posttest probabilities, PPV and NPV. We shall check that the use of the $LR^+$ and $LR^-$ from Schuh et al. gives us the same PPV and NPV as the direct use of sensitivity and specificity.

Using $LR^+ = 15.8$ and a pretest probability of 0.2, $Odds_{post}^+ = 15.8 \times 0.2/0.8 = 3.95$ then we can calculate $PPV = odds/(1 + odds) = 3.95/(1 + 3.95) = 0.798$, the same value we calculated above. Below, we use our R function PredictiveValuesLR to calculate PPV and NPV from prevalence and the $LR^+$ and $LR^-$ from the Schuh et al. data:

```
> PredictiveValuesLR(LRp = 15.8,LRn = 0.139, P = 0.2)
Pretest PPV NPV
0.2 0.798 0.967
```

### Nomograms

The formulas above allow accurate calculation of posttest probabilities from the relevant quantities. But, in many cases, the assessment of pretest probability gives only an approximate value and a similarly approximate value for the posttest probability will suffice for clinical decision making. The relationship between the pretest probability, the likelihood ratio, and the posttest probability can be represented in a nomogram, a graphical aid that is frequently used by physicians to obtain approximate posttest probabilities based on pretest probabilities and likelihood ratios (Fig. 3.1). To use the nomogram, draw a straight line from the pretest probability of disease on the left axis through the value corresponding to the likelihood ratio of the diagnostic test and over...
Inference about the LR usually focuses on estimation of confidence intervals, rather than hypothesis testing. When small samples and zero counts are not an issue, an approximate confidence interval can be calculated using the formula implemented in our R program `DiagTest`, which estimates all the measures we have discussed, along with their confidence intervals. For the data in Table 3.3, we can use the command

```r
> DiagTest(TP=86, FP=9, FN=13, TN=155)
```

<table>
<thead>
<tr>
<th>PointEst</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>86.900</td>
<td>78.8000</td>
</tr>
<tr>
<td>spec</td>
<td>94.500</td>
<td>89.9000</td>
</tr>
<tr>
<td>PPV</td>
<td>90.500</td>
<td>83.0000</td>
</tr>
<tr>
<td>NPV</td>
<td>92.300</td>
<td>87.2000</td>
</tr>
<tr>
<td>LRP</td>
<td>15.829</td>
<td>8.3489</td>
</tr>
<tr>
<td>LRn</td>
<td>0.139</td>
<td>0.0836</td>
</tr>
</tbody>
</table>

Inference about the LR' and LR' usually focuses on estimation of confidence intervals, rather than hypothesis testing. When small samples and zero counts are not an issue, an approximate confidence interval can be calculated using the formula implemented in our R program `DiagTest`, which estimates all the measures we have discussed, along with their confidence intervals. For the data in Table 3.3, we can use the command

```r
> DiagTest(TP=86, FP=9, FN=13, TN=155)
```

![Fig. 3.1](image) Nomogram showing conversion of pretest probability through a likelihood ratio to a posttest probability for LR' = 15.8 and LR' = 0.139.

to the right axis. The value where the line crosses the right axis is the posttest probability of disease. The left panel of **Fig. 3.1** shows a nomogram being used to calculate posttest probabilities for a patient who, after a clinical examination, had an 85% pretest probability of appendicitis. Suppose that the likelihood ratio is 15.800 for a positive test and 0.139 for a negative test. A line drawn from 85% on the pretest axis through a likelihood ratio of 15.8 shows that the posttest probability is almost 99% after a positive ultrasound and around 45% after a negative ultrasound. The right panel of **Fig. 3.1** shows the use of the nomogram for the same LRs when the pretest probability is 20%. Notice that the numerical value on the LR axis can usually be located only approximately and that the posttest probabilities can be read off only approximately; in many cases, the degree of approximation is good enough in a clinical setting. If more precision is desired, then the formulas should be used.
**Fig. 3.2** Posttest probability versus pretest probability for several landmark values of the positive and negative likelihood ratios.

**Fig. 3.2** shows how the posttest probability is altered by likelihood ratios of different sizes. If the LR is 1, or nondiagnostic, then the posttest probability will be equal to the pretest probability and fall on the diagonal line. The vertical distance between the diagonal line and the curve for the relevant LR is the absolute increase in the probability of disease associated with that LR. As an example, suppose the LR+ is 10 and the pretest probability of disease is estimated to be 20%, a positive diagnostic test increases the probability to a value of 71.4%. This positive predictive probability can be calculated with a call to our function using LRp = 10 and any value for LRn:

```r
> PredictiveValuesLR(LRp = 10, LRn=1, 0.2)
```

### Positive Likelihood Ratios for Multilevel Outcomes

Sensitivity and specificity are not defined for a multilevel diagnostic test, but techniques exist to help the researcher identify the
optimal level at which to split a multilevel diagnostic test to create a binary diagnostic test so that methods for binary diagnostic tests can be applied. If the multilevel diagnostic test is not going to be dichotomized, one can compute the likelihood ratio for each level of the test. In this instance, the LR is the relative increase in the pretest odds of disease for subjects obtaining that particular level.

Example: Low-Dose Abdominal CT for Assessment of Appendicitis

Kim et al.\(^9\) compared low-dose and standard dose abdominal CT to evaluate suspected appendicitis in a sample composed mainly of adults. Out of a total of five grades on the CT finding, they considered a grade of 3 or more to be a positive diagnosis. Table 3.6 displays the data for the low-dose CT group for this definition of a positive finding.

Using our R function DiagTest, we can easily calculate the diagnostic measures of interest when the five levels of the diagnostic test are grouped into the 2 × 2 table in Table 3.6.

> DiagTest(TP = 156,FN = 9,FP = 18,TN = 250)

<table>
<thead>
<tr>
<th>PointEst</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>94.5000</td>
<td>90.000</td>
</tr>
<tr>
<td>spec</td>
<td>93.3000</td>
<td>89.600</td>
</tr>
<tr>
<td>PPV</td>
<td>89.7000</td>
<td>84.200</td>
</tr>
<tr>
<td>NPV</td>
<td>96.5000</td>
<td>93.500</td>
</tr>
<tr>
<td>LRp</td>
<td>14.0768</td>
<td>8.997</td>
</tr>
<tr>
<td>LRn</td>
<td>0.0585</td>
<td>0.031</td>
</tr>
</tbody>
</table>

By most standards these diagnostic measures are very good, but it is natural to ask what information is lost by grouping patients with different results on CT. Are patients with a grade of 5 more likely to have appendicitis than patients with a grade of 3? If we accept that binary grouping is necessary because patients in the ER need to be either treated or sent home (a binary decision), then is a grade of 3 the optimal threshold for making this decision? We deal with the first question first.

Table 3.7 shows the numbers of patients with and without a confirmed diagnosis of optimal level at which to split a multilevel diagnostic test to create a binary diagnostic test so that methods for binary diagnostic tests can be applied. If the multilevel diagnostic test is not going to be dichotomized, one can compute the likelihood ratio for each level of the test. In this instance, the LR is the relative increase in the pretest odds of disease for subjects obtaining that particular level.

Example: Low-Dose Abdominal CT for Assessment of Appendicitis

Kim et al.\(^9\) compared low-dose and standard dose abdominal CT to evaluate suspected appendicitis in a sample composed mainly of adults. Out of a total of five grades on the CT finding, they considered a grade of 3 or more to be a positive diagnosis. Table 3.6 displays the data for the low-dose CT group for this definition of a positive finding.

Using our R function DiagTest, we can easily calculate the diagnostic measures of interest when the five levels of the diagnostic test are grouped into the 2 × 2 table in Table 3.6.

> DiagTest(TP = 156,FN = 9,FP = 18,TN = 250)

<table>
<thead>
<tr>
<th>PointEst</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>94.5000</td>
<td>90.000</td>
</tr>
<tr>
<td>spec</td>
<td>93.3000</td>
<td>89.600</td>
</tr>
<tr>
<td>PPV</td>
<td>89.7000</td>
<td>84.200</td>
</tr>
<tr>
<td>NPV</td>
<td>96.5000</td>
<td>93.500</td>
</tr>
<tr>
<td>LRp</td>
<td>14.0768</td>
<td>8.997</td>
</tr>
<tr>
<td>LRn</td>
<td>0.0585</td>
<td>0.031</td>
</tr>
</tbody>
</table>

By most standards these diagnostic measures are very good, but it is natural to ask what information is lost by grouping patients with different results on CT. Are patients with a grade of 5 more likely to have appendicitis than patients with a grade of 3? If we accept that binary grouping is necessary because patients in the ER need to be either treated or sent home (a binary decision), then is a grade of 3 the optimal threshold for making this decision? We deal with the first question first.

Table 3.7 shows the numbers of patients with and without a confirmed diagnosis of

<table>
<thead>
<tr>
<th>True state (D)</th>
<th>Appendicitis (D = 1)</th>
<th>No appendicitis (D = 0)</th>
<th>LR calculation</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>185</td>
<td>(2/165)/(185/268)</td>
<td>0.018</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>65</td>
<td>(7/165)/(65/268)</td>
<td>0.175</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>11</td>
<td>(13/165)/(11/268)</td>
<td>1.92</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>3</td>
<td>(53/165)/(3/268)</td>
<td>28.7</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>4</td>
<td>(90/165)/(4/268)</td>
<td>36.5</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>268</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
appendicitis falling into each of the five grades on low-dose CT. For each grade \( g \) on CT, the LR is calculated as the ratio of the proportions \( P(\text{Grade } = g \mid D = 1) / P(\text{Grade } = g \mid D = 0) \). The calculations are shown in the fourth column of Table 3.7 and the resulting LRs in the final column.

The observed LRs increase as CT grade increases. A grade of 3 confers only a doubling of pretest odds of appendicitis, but in Table 3.6, grade 3 is grouped with grades 4 and 5, which have LRs of around 30 or more. A grade of 3 is relatively inconclusive, and indeed, Kim et al.⁹ report the overall percentage receiving a grade of 3, labeling it indeterminate.

Our program LRCI can be used to obtain a confidence interval for the LRs of each of the CT grades. For grades 4 and 5, the LRs can be calculated as

\[
> \text{LRCI}(x1=53, n1=165, x2=3, n2=268)
\]

PointEst Lower Upper
28.69 9.11 90.34

\[
> \text{LRCI}(x1=90, n1=165, x2=4, n2=268)
\]

PointEst Lower Upper
36.5 13.7 97.6

These confidence intervals are wide, ranging from about a third of the LR to three times the LR, and overlap substantially. This is perhaps another reason to prefer grouping CT grade findings in this study: there are few patients with any particular individual values of low CT grades and appendicitis or high CT grades and no appendicitis.

We have seen that the cutoff of a CT grade of 3 can be justified by looking at the LR at each actual CT grade and then grouping all grades that increase the odds of appendicitis as a positive test and grouping all the grades that decrease the odds of appendicitis as a negative test. A technique that falls under the general title of ROC analysis can also be used to examine the diagnostic sensitivity and specificity for each potential dichotomous grouping and identify a grouping that provides a balance of sensitivity and specificity appropriate for the particular situation.

### Receiver Operating Characteristic Curves

Often, a test provides an ordinal or continuous value, which is then dichotomized and used to make a diagnosis. As one varies the cutoff, the true-positive and true-negative proportions will change. A higher cutoff leads to fewer positive diagnoses and fewer false-positives. Conversely, a lower cutoff results in more true-positive diagnoses and a larger number of false-positives. Fig. 3.3 illustrates this relationship with the data from Kim et al.⁹ Each of the four panels shows the same bar plots of the distribution of the CT grades for the patients with and without appendicitis, colored red when the grade exceeds the corresponding threshold and colored gray when it does not.

In the top left panel, where only a grade of 5 defines a positive CT, the red-colored cases represent only 54.5% of true cases; but only 1.2% of non-cases have a grade of 5, so specificity is 98.8%. Moving to the bottom left panel, the definition of a positive CT expands to include a grade of 4 or 5, and it can be seen that 32.2% more cases of appendicitis meet the criterion for a positive CT with only a minor increase of the false-positive fraction to 2.6%. When a positive CT is defined as a grade of 3, 4, or 5 (top right), sensitivity increases to 94.5% and the false-positive fraction increases to 5.5%. Finally, with a positive CT being defined as a grade of 2, 3, 4, or 5 (bottom right), the sensitivity has risen to 98.8% and the false-positive fraction has risen to 31%. Fig. 3.3 is presented for pedagogical purposes only as it illustrates the notion that if a diagnostic test has \( K \) levels (here \( K = 5 \)), there are \( K - 1 \) (here, 4) cutoffs that can each give a different sensitivity and false-positive fraction. In practice, any one of these four panels would suffice to show the distribution of CT grades in the two disease groups.

The four pairs of false-positive \((1 - \text{specificity})\) fraction and sensitivity values can be plotted with false-positive on the \( x \)-axis and sensitivity on the \( y \)-axis by convention. These are usually supplemented in the plot with two additional points: (a) at the bottom left, a point at false-positive fraction = 0 and sensitivity = 0, which
represents a decision to call everyone negative on the diagnostic test (or a hypothetical threshold for a positive test that is higher than any observed values, resulting in no positive diagnostic tests); and (b) at the upper right, a point at false-positive fraction = 1 and sensitivity = 1, which represents the rather useless threshold that calls everyone positive regardless of the test result (or equally, a threshold that is lower than the lowest observed value of the diagnostic test, so that all diagnostic test outcomes are positive). The resulting figure is called a receiver operating characteristic (ROC) plot.

To sum up, the underlying assumption of ROC curve analysis is that a diagnostic variable is used to discriminate between two
mutually exclusive states of tested individuals, diseased and nondiseased. The diagnostic sensitivity and specificity vary with changes in the selected cutoff values.

The R library \texttt{pROC}\textsuperscript{11} provides several tools for producing and plotting ROC curves. A one-time installation is needed before it is first used:

\begin{verbatim}
> install.packages("pROC")
\end{verbatim}

To make the routines in \texttt{pROC} available during an R session, the library must be attached using the command \texttt{library(pROC)}.

The data for the low-dose CT study are stored in a data file that can be read in with the command

\begin{verbatim}
ldct <- read.csv("ldcv.csv", header=T)
\end{verbatim}

The appendicitis variable is the binary disease state, with 1 representing appendicitis and 0 representing no appendicitis. The CT variable is the five-level CT grade. To check the distribution of CT scores in each group, use

\begin{verbatim}
with(ldct, table(CT, Appendicitis))
\end{verbatim}

To create an object representing an ROC curve based on exactly the process represented in Fig. 3.4, use the \texttt{roc} function.

\begin{verbatim}
roc1 <- with(ldct,
roc(response=Appendicitis,
predictor=CT))
\end{verbatim}

The ROC object can be plotted with the code below; the \texttt{type="o"} option overlays the sensitivity and false-positive points on the line that joins the points. By default, the \texttt{pROC} package labels the x-axis with specificity with the axis running from 1 to 0. We can obtain the more usual $1 - \text{specificity}$ (false-positive) axis with the option \texttt{legacy.axes=T}:

\begin{verbatim}
> plot(roc1, type="o", legacy.axes=T)
\end{verbatim}

A curve that coincides with the upper corner of the plot would perfectly discriminate between two patient groups, whereas a test that gives a diagonal line from the bottom left corner to the top right corner is useless in discriminating between two groups. It is unusual to see such an ROC curve. More commonly, the ROC curve allows the visualization of the trade-off between the true-positive and false-positive fractions for all relevant values of the threshold for dichotomization. The point closest to the top left corner (the one with the smallest value of $(1-\text{TP})^2 + \text{FP}^2$) represents one choice of a best cutoff value as it has both a high true-positive fraction and a low false-positive fraction. For the Kim et al.\textsuperscript{9} low-dose CT data, this point (A) corresponds to a positive CT being a grade of 3 or higher, as in the publication. Another choice of the optimal threshold results from maximizing the sum of sensitivity and specificity. These choices assume the “costs” of a false-positive and false-negative are the same. In Fig. 3.4, a marginal improvement in the sensitivity can be obtained at the expense of a large increase in the false-positive fraction by choosing point B instead of point A. If the consequences of missing cases of appendicitis are much higher than the consequences of wrongly classifying a healthy child as having appendicitis, this tradeoff may be worth it.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.4.png}
\caption{ROC curve for the data from Kim et al.\textsuperscript{9} showing (A) the point closest to the top left of the figure and (B) a point with higher sensitivity, at the expense of lower specificity.}
\end{figure}
While it is possible to formally assign costs to the two types of errors and compute an optimal threshold to minimize cost, it is more common to informally weigh the consequences of the two types of incorrect decision and use clinical judgment to choose a threshold.

If the ROC curve is used to identify an optimal threshold, keep in mind that the resulting sensitivity and specificity are likely overestimates of the true values. The threshold optimizes performance in this particular data set; with a new set of subjects, the values will generally be lower. Various solutions have been proposed to obtain unbiased estimates of diagnostic performance.

The area under the ROC curve (AUC or AUROC) is a single number summary of diagnostic accuracy across all the possible thresholds for a diagnostic test. The largest value for the AUC is 1 and this occurs when there is a threshold that gives complete separation of the diseased and nondiseased patients. All of the diseased patients are at or above the cutoff (so sensitivity is 100%) and all of the nondiseased patients are below it (so the false-positive fraction is 0). This point is at the top left corner of the ROC plot, so the curve is in effect a square with area 1. If every threshold results in a true-positive fraction that is equal to the false-positive fraction, then the ROC curve will be a diagonal line sloping from the bottom left to the top right and AUC = 0.5. The most useful interpretation of the AUC is that AUC is equal to the probability that a randomly chosen diseased person will have a higher diagnostic test result than a randomly chosen nondiseased person. It is not the probability that all diseased patients have higher scores than all nondiseased patients, so it can be quite high even when the diseased and nondiseased patients have overlapping distributions of the diagnostic test result. Another way to interpret the AUC is as the average sensitivity over a range of false-positive fractions. When comparing two different diagnostic tests, the one with the larger AUC has, on average, higher sensitivity for any given false-positive fraction. The AUC for the Kim et al. CT data and its 95% CI can be found by printing the `roc1` object created earlier.

```r
> print(roc1)

Data: CT in 268 controls (Appendicitis 0) < 165 cases (Appendicitis 1).

Area under the curve: 0.97
95% CI: 0.953-0.986 (DeLong)
```

The ROC curve analysis confirms that low-dose CT is an accurate diagnostic test for appendicitis and also supports the choice of Kim et al. of grades of 3 or higher as a positive diagnostic finding.

As a second example of ROC curve analysis, we will look at a portion of the data in Pisano et al., which compared digital and film mammography for breast cancer screening. Table 3 in that paper summarizes the number of women screened who were assigned malignancy scores of 1 through 7 and the number of cancers that were identified in each of those groups. This process was repeated for scores from digital and from film mammography. The dataset Pisano.csv uses this data to re-create the diagnostic test findings for each of the film and the digital mammograms in the full cohort of over 42,000 women. Without having access to the actual data though, it is not possible to re-create the data set that contains the paired digital and film mammograms on each woman. R code found on the accompanying website reads in the data and produces ROC curves for each of the two types of mammogram.

The resulting plots show the sensitivity and false-positive fraction from the approach outlined above—calculating the values by dichotomizing at all possible thresholds. These have solid points and straight lines joining them. The smaller AUC in the graph titles is the value calculated by adding up the area under the triangles and trapezoids between the x-axis and these straight lines. The curved lines are from a model-based approach that assumes the seven categories of malignancy score arise from breaking up an underlying continuous distribution at six breakpoints. The particular distribution assumed in this analysis is the...
normal distribution: the assumption is that the underlying “risk” is normally distributed in the diseased and nondiseased groups, with a different mean and a different standard deviation for the two groups. Since the normal distribution is continuous, there are an infinite number of thresholds, each giving different sensitivity and false-positive fractions. The smooth curves on Fig. 3.5 represent the sensitivity and false-positive fractions from this underlying model, called the binormal model because it is based on two normal distributions. It is usually the case that the binormal model ROC curve will have a larger AUC than the simpler nonparametric ROC curve. Here, the AUC is about 0.05 higher when estimated by the binormal model and we see that the smoothed curve is far above the line segment joining the point with highest sensitivity to the top right corner. There is no threshold that generates data in this region of the graph (with high sensitivity and a moderate to high false-positive fraction) to assess whether the model is correct here, but under the binormal model, the sharp rise of the curves on the left means that they must follow that path to the top right corner.

More recently, ROC curve methods have been applied to continuous test results. Several methods exist to create an ROC plot based on continuous data. The simplest method involves plotting pairs of true-positive rates (sensitivity, y-axis) versus false-positive rates (1−specificity, x-axis) at all possible values for the test cutoff, where the empirical distributions (essentially the histograms) for the diseased and healthy patients are calculated. This method is usually called empirical or nonparametric since the distributions for the two groups are unstructured. This approach is robust since it is free of any structural assumptions. A disadvantage is the jaggedness of the resulting ROC plot. If one uses an alternative approach assuming a parametric model of the empirical sample, an ROC curve can be fitted. The binormal model, assumed to underlie the seven categories in the Pisano example above, can also be applied to continuous outcomes. Empirical evaluation of the performance of this model in the face of departures from binormality has been limited to interpretations of radiology-type examinations recorded on rating and interval scales.

Comparing Results from Two Diagnostic Tests

Often, an investigator wishes to compare two diagnostic tests. There are several aspects
of a diagnostic test that can be compared: sensitivity and specificity for binary diagnostic tests or the AUC for diagnostic tests taking on several levels. When a sensitivity or specificity is being compared between two different samples of patients, this is simply a two-sample test of proportions (Chapter 13). This approach can be used to obtain the difference in sensitivity and its 95% confidence interval, as well as a $p$ value for a test of the null hypothesis of equal sensitivities.

**Low-Dose Abdominal CT for Assessment of Appendicitis Example**

In the study of low-dose CT in Kim et al., the primary interest of the study was to determine whether low-dose CT performs as well as standard-dose CT when diagnosing appendicitis. Earlier, the sensitivity from the low-dose group was found to be $156/165 = 94.5\%$ and for the standard-dose group it was found to be $171/180 = 95.0\%$. The difference of the sensitivity between the two types of tests is $0.945 - 0.950 = -0.005$ or $0.5\%$. The 95% confidence interval can be found using the R function `prop.test`

```r
> prop.test(x=c(156,171),
n=c(165,180),correct=F)
```

- data: c(156, 171) out of c(165, 180)
- X-squared = 0.035967, df = 1,
- p-value = 0.8496

alternative hypothesis: two.sided
95 percent confidence interval:
-0.05160238 0.04251147
sample estimates:
prop 1 prop 2
0.9454545 0.9500000

The difference in sensitivity values has a 95% confidence interval of $(-0.052, 0.042)$ or $(-5.2\%, 4.2\%)$. Since 0 is in the interval, there is no statistical difference in the sensitivities between the two tests at an $\alpha$ of 0.05. Using this function with the specificity numbers, we see the difference between the specificities of the two tests is $0.933 - 0.938 = -0.0056$ and the 95% confidence interval for the difference is $(-0.048, 0.036)$. The reader may notice the confidence intervals reported in Kim et al. are slightly different. This is due to slightly different but more complex methods that are used to construct the confidence interval. Again, 0 is contained in this second 95% confidence interval indicating no statistically significant difference in the specificity between the two groups.

Note, in the text here we have considered comparing tests between two different samples. In particular, no one subject got two tests; rather, each test was performed on two independent groups of patients. If one is interested in comparing two tests using the same set of patients (i.e., each patient gets both tests), the above-mentioned methods are not appropriate. Methods are required to address the correlation within each patient (paired methods) and one can find more details in Martincich et al. provide an example using paired methods that compare two different types of contrast agent in breast MRI. A common design that can be analyzed with basic methods is one that classifies each subject on two binary diagnostic tests as well as the reference standard. In this design, it is common to ask whether the two diagnostic tests have equal sensitivity or equal specificity. Atri et al. conducted a study where a staff radiologist, a fellow, and a resident used multidetector CT to identify surgically important bowel or mesenteric injury in patients who had experienced blunt trauma. Each set of images was assessed by each of the three readers and the reference standard was a laparatomy-confirmed finding of injury or no injury. Here, we focus on the comparison of the diagnostic accuracy of the staff radiologist and the fellow. To compare sensitivity of interpretation of CT findings by these two readers, we need to identify the cases with surgically important injury and compare their findings on these cases. There are four possible combinations of outcomes: (1) both readers find the case
answer by asking the question in a slightly different way and focusing on the 7 + 3 cases where the two readers disagreed. We shall assess whether these disagreements are equally in favor of the staff radiologist being correct (giving a positive assessment) and the fellow being correct. That is, given there are 10 disagreements, and 7 in favor of the fellow, we shall test the hypothesis that there is a 50:50 chance that it is either the fellow or the staff that is correct. Unlike the McNemar test, this approach does not rely on the sample being large to calculate the correct p value.

\[ \text{binom.test(7,10)} \]

Exact binomial test
data: 7 and 10
number of successes = 7, number of trials = 10, p-value = 0.3438
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.348 0.933

We can conclude that the interpretations of the staff radiologist and of the fellow do not lead to statistically significant different sensitivities of CT to identify surgically important bowel injury. When we compare specificities, we find that the staff radiologist correctly identifies \((36 + 16) / 59\) cases as not having significant injury whereas the fellow identifies \((36 + 3)\) such cases. The staff

<table>
<thead>
<tr>
<th>True bowel injury</th>
<th>Fellow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Positive</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No bowel injury</th>
<th>Fellow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>Positive</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>20</td>
</tr>
</tbody>
</table>

Note: The left table shows results for those with a true bowel injury and the right table shows the results for those with no bowel injury.

*Staff radiologist and a fellow.

Table 3.8  Findings for two readers* examining the same multidetector CT images

<table>
<thead>
<tr>
<th>True bowel injury</th>
<th>Fellow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Positive</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No bowel injury</th>
<th>Fellow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>Positive</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>20</td>
</tr>
</tbody>
</table>

The difference in the proportions is not statistically significant. We can get the same

We can conclude that the interpretations of the staff radiologist and of the fellow do not lead to statistically significant different sensitivities of CT to identify surgically important bowel injury. When we compare specificities, we find that the staff radiologist correctly identifies \((36 + 16) / 59\) cases as not having significant injury whereas the fellow identifies \((36 + 3)\) such cases. The staff
radiologist’s interpretation has a higher observed specificity (88%) than the fellow (66%). If we examine the 19 cases where the two readers disagree, we find that the staff radiologist was correct in 16 of them and the fellow was correct in 3.

> binom.test(16, 19)

gives a p value of 0.004 for a test of the true probability being 0.5. We reject the hypothesis that when the two readers disagree on the truly negative cases, they have an equal chance of being correct and conclude that the staff radiologist has higher specificity for identifying this condition on multifocal CT.

When two or more multilevel diagnostic tests have been assessed on the same subjects, methods exist\(^{19,20}\) that compare the AUC between tests, accounting for the correlation. The \texttt{roc.test} function in the \texttt{pROC} package in R implements a number of tests of paired ROC data.

### Case Studies

#### Case 1

Typically, on MRI, fat has short T1 and long T2 relaxation times and is bright on the T1-weighted image and gray on the T2-weighted image. On the other hand, fluid has long T1 and T2 relaxation times and is dark on the T1-weighted image and bright on the T2-weighted image. A 12-year-old boy had pain in his right shoulder for a couple of months (Fig. 3.6). If an investigator wants to evaluate

---

**Fig. 3.6** Example of sensitivity of MR imaging for diagnosis of a fat-containing tumor. (a) Frontal radiographic view, and (b) unenhanced axial T1- and (c) fat-saturated contrast-enhanced axial T1-weighted MR images of a lesion in the proximal right humerus.
the false-negative rate of MRI for diagnosing bone tumors, then it will be important to design a diagnostic test research study that gives a precise estimate of the sensitivity of MRI.

**Findings**

The biopsy of this tumor revealed a pleomorphic liposarcoma. Based on the MRI result alone (without prior knowledge of the histopathologic result), one would not consider this as a fatty tumor since it has predominant low signal on T1-weighted images. This case would be a false-negative result. This tumor often has heterogeneous/mixed low, intermediate, and/or high signal intensity on T1-weighted images. Less than 26% of pleomorphic liposarcomas have fat signal on MRI.\(^{21,22}\)

**Case 2**

An 8-year-old boy previously diagnosed with grade IV neuroblastoma and presenting with known bone metastases had new onset of left elbow swelling and pain. **Fig. 3.7** shows bony and soft tissue involvement of the mid and lower aspects of the left humerus on MRI. The \(^{111}\)In white blood cell (WBC) nuclear medicine scan demonstrates increased uptake of the radiopharmaceutical agent in the left elbow. Results of the metaiodobenzylguanidine (MIBG) nuclear medicine scan show increased uptake in the region of the left elbow suggesting metastatic disease. The research question is whether \(^{111}\)In WBC is specific for diagnosing superimposed infection in a patient with bone metastases from...

---

**Fig. 3.7** Example of specificity of 111-Indium white blood cell test for diagnosis of superimposed infection in a patient with bone metastases. *(a)* Postcontrast sagittal T1-weighted image with fat saturation, *(b)* nuclear medicine scan, and *(c)* metaiodobenzylguanidine (MIBG) nuclear medicine scan.
known neuroblastoma; that is, does it have a low false-positive rate?

Findings

The $^{111}$In WBC nuclear medicine scan demonstrates increased uptake of the radiopharmaceutical agent in the left elbow and is specific with regard to the etiopathogenesis of the process (could be caused by an acute infection). Results of the MIBG nuclear medicine scan, which is specific for neuroectodermal tumors, show an increased uptake in the region of the left elbow suggesting metastatic disease. This is a case of grade IV neuroblastoma with diffuse bone metastases. The lesion in the distal left humerus represents infection superimposed on metastatic tumor (WBC and MIBG confirmation).

Case 3

Ultrasound can be used to visualize the adult brain through an area of the skull called the temporal bone window, just in front of and above the ear. Through this relatively thin part of the skull ultrasound can be used to visualize the brainstem, which is one of the sites of neurodegeneration in Parkinson’s disease. An area of hyperechogenicity that appears to correspond to the substantia nigra within the midbrain can be seen as a hyperechogenic focus in a proportion of individuals with and without Parkinson’s disease. It has been well documented that the cross-sectional area of this hyperechogenicity is, on average, larger in patients with Parkinson’s disease than in those without the disease. Of greater interest, this area of hyperechogenicity does not increase with disease duration and is enlarged even from very early on in the disease. This has raised the possibility that transcranial ultrasound could be used as a predictive test for the future development of Parkinson’s disease. To test this hypothesis, Berg and colleagues conducted a prospective cohort study on 1847 older adults free of Parkinson’s disease at baseline. A total of 1400 had both transcranial sonography at baseline and follow-up, averaging 3 years. Scans were classified as hyperechogenic if the cross-sectional area of hyperechogenicity exceeded the 90th percentile of scans from that region. Over the period of follow-up 10 participants developed Parkinson’s disease and 8 of them (sensitivity = 80%) had hyperechogenic scans at baseline. Among the 1390 individuals who did not develop Parkinson’s disease on follow-up, 1136 (specificity = 82%) did not have hyperechogenic scans at baseline. These results confirm the study hypothesis but the question remains whether transcranial ultrasound is a useful predictive test for the future development of Parkinson’s disease.

The positive predictive value of substantia nigra hyperechogenicity can be calculated from the number of true-positive cases among all cases with a positive ultrasound: PPV = a / (a + b) = 8 / (8 + 254) = 0.03. Since the sample chosen for this study comprises older adults without Parkinson’s disease, this is a valid estimate of the PPV. The pretest probability was increased from 10/1400 = 0.0071 to a posttest probability of 0.03, an increase that is also seen in LR+ of sensitivity/(1 − specificity) = 0.80/0.183 = 4.37. Although there is a higher likelihood of a positive test in those who develop Parkinson’s disease than in those that don’t (LR+ = 4.37), the low prevalence of the disease means that the positive predictive value of the test is very low. Therefore, for at least over 3 years of follow-up, the proposed ultrasound finding is not useful for predicting the development of Parkinson’s disease. Interestingly, the sensitivity (80%) and the specificity (82%) are both quite high. This dramatic difference between specificity/sensitivity and positive predictive value (3%) results from the very low prevalence (or in this case incidence) of the disease in the population tested.

Bias in Diagnostic Tests

Bias is a mechanism that produces lack of internal validity or incorrect assessment of the association between, in this setting, a diagnostic test and its reference standard. In statistical terminology, if there is bias, the expected values of the estimated sensitivity or specificity are not equal to their true values in the target population. Biases can be classified by the...
Faster-growing tumors generally have a shorter asymptomatic phase than slower-growing tumors, and so are less likely to be detected. However, faster-growing tumors are also often associated with a poorer prognosis. Slower-growing tumors are hence likely to be over-represented in screening tests. This means that screening tests can be erroneously associated with improved survival, even if they have no actual effect on prognosis from the true onset (not detectable onset) of disease.25,26 This phenomenon is illustrated in Fig. 3.9.

- **Lead time bias**: This occurs during the period of time between the detection of the medical condition by screening and the period of time when it would ordinarily be diagnosed clinically (after presentation of symptoms).3,25 As an example, two tests for diagnosis of cancer are compared: one test (the new, experimental one) diagnoses the disease earlier, but there is no effect on the outcome of the disease. It may appear that this test prolonged survival, when in fact it only resulted in earlier diagnosis when compared to the traditional test that was conducted by the time symptoms were present. This bias should be considered during the evaluation of the effectiveness of a specific test (Fig. 3.8).

- **Length time bias**: This bias occurs because the proportion of slow-growing lesions diagnosed during screening programs is greater than the proportion of those diagnosed during usual medical care. It can affect data on screening tests for cancer. Faster-growing tumors generally have a shorter asymptomatic phase than slower-growing tumors, and so are less likely to be detected. However, faster-growing tumors are also often associated with a poorer prognosis. Slower-growing tumors are hence likely to be over-represented in screening tests. This means that screening tests can be erroneously associated with improved survival, even if they have no actual effect on prognosis from the true onset (not detectable onset) of disease.25,26 This phenomenon is illustrated in Fig. 3.9.

**Fig. 3.8** Lead time bias for diagnosis of cancer when two diagnostic tests (a new one that appears effective to improve survival time and a traditional one that appears ineffective to improve survival time) are compared.

- **Verification bias**: This occurs when the decision to perform the reference standard test to establish the presence or absence of disease is influenced by the results of the diagnostic test, along with other measured, or not measured, risk factors. If only data from patients who received the reference standard test are used to assess the test performance, the commonly used measures of diagnostic test performance—sensitivity and specificity—are likely to be biased. Sensitivity would often be higher, and specificity would be lower, than the true

- **Verification bias**: This occurs when the decision to perform the reference standard test to establish the presence or absence of disease is influenced by the results of the diagnostic test, along with other measured, or not measured, risk factors. If only data from patients who received the reference standard test are used to assess the test performance, the commonly used measures of diagnostic test performance—sensitivity and specificity—are likely to be biased. Sensitivity would often be higher, and specificity would be lower, than the true
values. This bias is called verification bias. This type of bias is also known as "workup" or referral" bias. Example: In a research study of diagnostic performance of ultrasound and CT for diagnosis of appendicitis, only patients who have a high likelihood of disease based on ultrasound or CT undergo surgery (the reference standard).

- **Ascertainment bias**: This is a type of sampling (selection) bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. For example, structural MRI assessment of cerebral atrophy is a late feature in the progression of the disease and only detects frank dementia. Therefore, it lacks signs of dementia in high functioning individuals in the beginning stages of the disease. If MRI is the diagnostic test used for detection of dementia in a research study, it will be positive for patients with late disease, but it will miss (false-negative result) patients with early disease.

- **Imperfect reference standard**: When there is no available reference standard for the true disease state, results of a new diagnostic test are often compared to those of other routinely used diagnostic procedures. If these routinely used tests do not classify a patient’s true disease status correctly (i.e., they themselves have less than 100% sensitivity and specificity against the “truth”), then it is possible that the sensitivity and specificity of the new test will be underestimated. This will happen if the new test is better than the old test.

### Highlights of Key Points

- The sensitivity \( \frac{a}{a + c} \) and specificity \( \frac{d}{b + d} \) are properties of a diagnostic test that indicate the probability that the test is positive in diseased subjects and negative in nondiseased subjects. Sensitivity and specificity do not indicate the probability that a subject has the disease in question. They do not depend on the prevalence of disease.

- The positive \( \frac{a}{a + b} \) and negative \( \frac{d}{c + d} \) predictive values from a study indicate the probability of disease after a positive or negative test in that study but they are valid only for the prevalence of disease in that study. The predictive values vary considerably with prevalence of disease for a given sensitivity and specificity.

---

**Fig. 3.9** Length time bias for diagnosis of cancer. (a) Slowly progressive tumors are more likely to be detected by the diagnostic test than rapidly progressive tumors. (b) Screening captured cancer in only 5 out of 12 (42%) cases of the disease, but most of these cases (4 out of 5, 80%) survived after 5 years of diagnosis. Out of the cases that were not diagnosed through screening, 5 out of 12 (42%) survived. Therefore, the most aggressive cases of cancer were not diagnosed by screening.
The likelihood ratio is a useful tool as it indicates how the odds of a disease change with a positive or negative test. The likelihood ratio from a study can be combined with a pretest probability for a new setting to produce a posttest probability for this new setting.

Positive likelihood ratio = sensitivity / (1 − specificity)

Negative likelihood ratio = (1 − sensitivity) / specificity

The posttest odds is equal to pretest odds × LR, using LR+ after a positive test and LR− after a negative test.

The receiver operating characteristic (ROC) curve shows the tradeoff between sensitivity and specificity for a diagnostic test with two or more potential thresholds for dividing cases into positive and negative diagnoses. The area under the ROC curve (AUC) is an overall measure of how well the diagnostic test separates the diseased and nondiseased subjects. This area is between 0.5 (no separation between diseased and nondisease subjects) and 1 (perfect separation).

References

21. Stephen PM. Tumor and tumor-like lesions of the superficial soft tissues including subcutaneous fat.
Learning Objectives

- To explain the importance of measurement properties to health outcomes research.
- To explain types of variables and scaling.
- To explain the terms validity, reliability, and responsiveness and to be able to describe these properties for a given measure.
- To explain the different components of validity (face, content, construct, and criterion validity) in the context of measurements.
- To describe sources of variation in measurements.
- To differentiate between different types of measurements and to create the appropriate research designs using these concepts.

Concepts

Measurements

Measurements is the process of systematically assigning numbers or words to physical and behavioral characteristics so that they can be defined, quantified, and differentiated. This approach is simple for objective measures such as height and age but becomes much more complex when we try to measure health-related clinical outcomes that impact patients. Most of these outcomes are actually very complex constructs, made up of many different components. If we consider an outcome as simple as the measurement of the symptom of pain, for example, we soon realize that it is difficult to describe “pain” using the categories or numbers that are essential for statistical analysis required in research. If investigators are interested in developing a research design for an intervention that modified pain, for example, they need an a priori definition of pain. In the context of a research protocol, this is done by assigning values to objects. A classic example of this process is how Wilson and Cleary first defined the concept of health-related quality of life (HRQoL) in 1995 by developing a classification scheme for different measures of health outcomes. They defined five aspects of health outcomes: biological and physiological factors, symptoms, functioning, general health perceptions, and overall quality of life. They also proposed causal relationships between them. Through this classification scheme researchers could include all these factors when measuring HRQoL and could focus on the aspects of HRQoL that were most relevant to their work. The result of this process has been the development of HRQoL questionnaires made up of a series of questions that include the five aspects of the construct being measured, resulting in summary scores that can be used for analysis.

The actual development of these specific measurement tools is very laborious and requires years of selecting and testing the items that should and should not be included in a scale that will measure as closely as possible the concept being measured. Before a measurement tool can be considered usable as a research outcome in future research, it must have been demonstrated to be reliable, valid (Fig. 4.1), and responsive. The development of a magnetic resonance imaging (MRI) scoring system for evaluation of knees and ankles of hemophilic children is one example of this process from the radiology literature. Once a given tool or scale or scoring system has been developed, it is specific only for the target population being tested. If one intends
Measurements: Validity, Reliability, and Responsiveness

be much greater. Similarly, in the example of an MRI scoring system for assessment of hemophilic arthropathy, the range of soft tissue and osteochondral tissue scores would be expected to be shifted toward more severe stages in places where prophylaxis to joint bleeding is not available to the population and toward less severe stages in places where prophylaxis is offered to the population. It is important that investigators consider the widest possible range of values for individual items of a scale or questionnaire to enable its application in different research settings, conditions, and health care systems.

When an investigator is designing his/her own research question, it is important to review the literature as to whether the outcome of interest has been previously measured with a validated measurement tool. Hopefully there will be a tool that approximates what the investigators intend to measure, and if luck is on the investigator’s side, there may be several, and this will allow him/her to choose the tool that most closely matches the target population and the research question.

When we state that a measurement tool is reliable, valid, and responsive, what we are really asking is, “How close is the measure from the truth?” The measurement properties to use the same MRI scoring system for healthy adults, for example, he/she should first demonstrate that the measurement tool is reliable, valid, and responsive for assessment of healthy adults. In the HRQoL example, the Short Form (SF)-36 is an example of a questionnaire that assesses quality of life in the general population, while the European Organization for Research and Treatment of Cancer Quality of Life (EORTC-QoL) measure is one that is specifically designed for cancer patients. Whereas one assumes that the same five components of the construct of HRQoL (biological and physiological factors, symptoms, functioning, general health perceptions, and overall quality of life) will make up the two questionnaires, SF-36 and EORTC-QoL, the actual items of each questionnaire would need to be different in the two populations of interest. The prevalence of symptoms of nausea and vomiting, for example, would be much more relevant for cancer patients as compared to the average population. In addition to the actual questions, the scale of the items of each questionnaire may need to change. Most healthy people, for example, would score around the same for pain (assuming not much pain in daily life), whereas for cancer patients the range of pain scores would be expected to

![Fig. 4.1](image-url)
of reliability (reproducibility of the measure when repeated multiple times), validity (the degree that the tool measures what it was supposed to measure), and responsiveness (the extent to which an instrument can measure change when change has occurred) are all different aspects trying to quantify and describe how a given measurement tool performs in relation to the truth. Fig. 4.1 illustrates the concepts of reliability and validity in relation to the truth.

### Types of Data

Measurements of clinical phenomena yield four main types of data: nominal, ordinal, interval, and continuous.6

**Nominal data** does not have meaning beyond representing unordered categories despite numbers often being used to represent the categories, such as, languages (German, French, English, Spanish), blood types (A, B, O, AB), sex (female, male), or identifiers of patients in a study. The categories are mutually exclusive and every subject fits into a category.

When the order among the categories is reflective of more or less of a phenomenon, the data are *ordinal*. Some examples include a pain scale ranging from 0 to 10 where 0 represents no pain and 10, writhing agony; the Arnold-Hilgartner radiographic scale for assessment of severity of hemophilic arthropathy7 where grade 1 represents mild changes and grade 5, late changes (Fig. 4.2); and synovitis domain of the Outcome Measures of Rheumatoid Arthritis Clinical Trials (OMERACT) ranging from 0 to 3 representing none, mild, moderate, and severe changes, respectively.8 Ordinal scales record information about the rank order of levels or scores. The distance between these levels is not necessarily equivalent even though the levels

---

**Arnold-Hilgartner plain radiograph grading system for hemophilic arthropathy:**

- **grade 0**: normal joint
- **grade I**: no skeletal abnormalities, soft-tissue swelling is present
- **grade II**: osteoporosis and overgrowth of the epiphysis, no cysts, no narrowing of the cartilage space
- **grade III**: early subchondral bone cysts, squaring of the patella, widened notch of the distal femur or humerus, preservation of the cartilage space
- **grade IV**: findings of grade III, but more advanced; narrowed cartilage space
- **grade V**: fibrous joint contractures, loss of the joint cartilage space, extensive enlargement of the epiphyses with substantial disorganization of the joint

---

Fig. 4.2 Examples of ordinal data whereby data is rank-ordered but the distance between each level is not fixed. (a) A commonly used Likert-type pain scale; (b) the Arnold-Hilgartner plain radiograph scale for assessment of severity of hemophilic arthropathy.9
are one unit apart. Consequently, arithmetic operations do not make sense for these data, and absolute values have little meaning. Despite this, many patient-based questionnaires sum up the responses of these ordinal data to create a summary score that is then considered to be interval level data that can be averaged, added, or subtracted to measure change.

*Interval or discrete data* are ordered and have magnitude that is meaningful, such as the number of motor vehicle accidents; height above the sea level; Fahrenheit and Celsius temperatures; dates of imaging examinations; posttest/pretest score differences in a course (educational research design); and radiographic scale for assessment of severity of hemophilic arthropathy based on counting the number of endorsed items (*Fig. 4.3*). There are equal distances between units or scores; therefore, addition and subtraction operations are acceptable with interval data. Scores are not related to a true zero, and can be negative. Zero simply represents an additional point of measurement.

*Ratio or continuous data* has a true zero and has an unrestricted range in that any number of decimal points may be used depending on the error of measurement, such as, length; weight; quantitative diagnostic tests: signal uptake in time-intensity curves (dynamic contrast-enhanced MRI) (*Fig. 4.4*); clinical tests such as surrogate measures of exposure (e.g., a series of questions to determine pesticide exposures); and outcomes. Addition, subtraction, multiplication, and division can be applied to ratio data.

---

### Plain radiograph grading system for hemophilic arthropathy

<table>
<thead>
<tr>
<th>Type of change</th>
<th>Findings</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osteoporosis</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Enlarged epiphysis</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Irregular subchondral surface</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Closed physis</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Joint space narrowing</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Subchondral cysts</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td>Erosions of joint margins</td>
<td>Absent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>1</td>
</tr>
</tbody>
</table>

**Maximum score: 7**

---

*Fig. 4.3* Interval data whereby the levels are rank-ordered and the distance between each fixed, but not related to a true zero. (a) A thermometer that measures temperature in degrees Fahrenheit or Celsius between −20 degrees and 40 degrees. (b) The final score of the radiographic scale for assessment of severity of hemophilic arthropathy. The “distance” between successive scores is equal: a score of 2 represents the presence of 2 of the items of the scale regardless of the type of finding. The final score of severity of arthropathy is based on how many positive findings are present, considering similar importance for each finding.
describe a behavior or characteristic. In each of these situations we must separate the truth versus the possible measurement error for these data. This is also referred to as signal and noise. The signal represents differences we are interested in detecting and noise is the error in the measurement over these differences. Measurement error is a threat to the validity of the data, that is, if the data is truly measuring/representing what it is purporting to measure.

In terms of a signal, validity represents the ability to detect the right signal from the data. Measurement error can have many sources.

Consider the situation in which the investigators want to evaluate the effectiveness of a new drug (intervention) in patients with osteoporosis. One eligibility criterion is that the subject has obtained baseline (prior to start of treatment) dual energy X-ray absorptiometry (DEXA) measurements.

What are the possible sources of measurement error in a DEXA scan?

What are the possible sources of error in the intervention?
59

4 Measurements: Validity, Reliability, and Responsiveness

Instrument variability is the variability in the measurement due to changing environmental factors such as set up of the machine, temperature, or different reagent lots.

Some variability of measurements can be attributed to random error (affecting reliability), and some can be attributed to systematic errors such as bias or confounding variables (affecting validity) (Fig. 4.7). The amount of variability will be determined by how abstract the outcome measure is. Some measures, such as measurement of bone mineral density from DEXA, can be expected to have much less variability than a quality of life or symptom measurement when repeated. This is because in the first case once the DEXA scanner is calibrated (as part of a daily approach for quality

Subject variability is the intrinsic biologic variability in the characteristic being measured, such as body weight, bone density, or mood.

Observer variability is due to the observer, and can result from differences in years of clinical experience, skill in using an instrument, or selection of words in an interview.

Suppose that a secondary outcome in this study is a quality of life questionnaire. What are the possible sources of error of measurement for this outcome?

Most measurements are subjected to three potential sources of error: those related to the subject (examinee), the examiner or observer, and the instrument (examination) (Fig. 4.6).

Subject variability is the intrinsic biologic variability in the characteristic being measured, such as body weight, bone density, or mood.

Observer variability is due to the observer, and can result from differences in years of clinical experience, skill in using an instrument, or selection of words in an interview.
Fig. 4.6 Framework for determining sources of measurement error. Strategies to reduce these sources of error are listed within each box.

Fig. 4.7 Types of variability of measurements: Random error, which affects reliability of results, and systematic error (bias), which affects validity of results. Example of a blood oxygen level dependent (BOLD) MRI experiment whereby variability is noted between measurements of pulse oximeter (test) and arterial gasometry (reference standard).
control thus minimizing any systematic deviation) and the region-of-interest is determined, a machine provides measurements that should be consistent. Repeated measures in a quality of life or symptoms instrument are much more dependent on subjective variation. Nevertheless, DEXA measurements can also be subjected to random error from minor deviations in positioning of the patient despite strict protocols to minimize the resulting error. In terms of magnitude of the random error, however, a major source of error in the quality of life questionnaire is the context: whether the questionnaire is completed in the clinic or at home, or whether it is a mailed vs. interviewer-administered questionnaire. Some data suggest that higher levels of disability are reported in mailed questionnaires than when interviews are administered, even when the interviewer is impartial.\(^9\)

Systematic errors can occur for both single- and multi-source measurements. In the latter case, if measurements of either the test or reference standard or both are incorrect, the (incorrect) difference in measurements between the reference standard and the test will persist if repeated measurements are done, thus resulting in a measurement bias (Fig. 4.7). Different types of measurement bias are described below as related to subject, instrument, or observer.

**Subject bias** is how the subject interprets the measurement based on his/her experiences; it may be conscious or unconscious. Example: Recall bias which is caused by differences in the accuracy or completeness of the recollections retrieved (“recalled”) by study participants regarding events or experiences from the past.

**Instrument bias** is a fault in the instrument itself, in that it does not accurately capture the construct being studied. Example: An instrument needing recalibration and therefore providing wrong results in a systematic way.

**Observer bias** is how the observer interprets the results of the instrument due to his/her own conscious or unconscious expectations. Example: If the observer knows the goals of the study or the hypotheses, he/she may allow this knowledge to influence their observations during the study.

### How to Control for Measurement Errors?

Strategies to reduce measurement bias include:

1. **Standardize measurement methods:**
   - Educate subjects to follow instructions before assessment (e.g., to drink the amount of oral contrast prescribed prior to a CT scan of the abdomen)—to minimize subject variability
   - Mandate technique (e.g., standardize imaging protocols in an operations’ manual)—to minimize observer and instrument variability
   - Calibrate the instrument if appropriate—to minimize instrument variability

2. **Train operators** (e.g., for ultrasound scanning)—to minimize observer variability

3. **Introduce items to assess performance in the middle of the scale**—to minimize observer variability

4. **Automate outcome measurement** (e.g., for data analysis of regions of interest of contrast-enhanced MRI, use a priori determined boxes that are copied and pasted in predefined regions of the images, if appropriate, in all examinations)—to minimize observer variability

5. **Eliminate (and explain the cause of) outlier observers**—to minimize observer variability

6. **Blinding:** Blinding is a highly effective strategy used in observational and experimental research to avoid investigators and participants in a trial changing their behavior or reporting of outcomes in a systematic way (i.e., be biased) if they are aware of which patients receive which treatment, procedure, or diagnostic test.\(^{10}\) Whenever possible, researchers should be unaware of the treatment, procedure, or diagnostic test groups to which patients have been selected so that this knowledge cannot cause them to act differently, thus avoiding
Therefore, the greater the error, the less reliable the instrument will be. By recognizing the three aforementioned sources of error (subject, observer, and instrument), there are a number of strategies that can be used to improve reliability, depending on the instrument (Fig. 4.6). There are several different ways that reliability of an instrument can be described and tested, by describing its internal and external reliability.

**Internal Reliability**

Internal reliability, or consistency, evaluates the relationship of the test items in a multi-item scale to each other in order to ensure that all test items are measuring the same concept. Cronbach’s α (alpha) provides a measure of the internal consistency of a test or scale; it is expressed as a number between 0 and 1. If the items in a test are correlated between each other, the value of alpha is increased. However, a high coefficient alpha does not always mean a high degree of internal consistency. This is because alpha is also affected by the number of items in the test. If the test length is too short, the value of alpha is reduced. Thus, to increase alpha, more related items testing the same concept should be added to the test. A multi-item scale is generally considered internally reliable if a Cronbach’s α (alpha) coefficient is >0.70 for group comparisons. First used by Lee Cronbach in 1951 for psychology questionnaires, Cronbach’s alpha statistic is widely used in health science research, business, nursing, and other areas whenever the question is raised as to what extent a series of items measure the same thing. Interestingly, Cronbach called it the alpha coefficient because he intended to describe further coefficients. It is defined as:

\[ \alpha = \frac{K}{K - 1} \left( 1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2} \right) \]

where X is the sum of K components (items, questions) such that X = Y1 + Y2 + Y3 + Yk. Here, \( \sigma_X^2 \) is the variance of the observed total scores, and \( \sigma_{Y_i}^2 \) is the variance of the component i for the current sample. Cronbach’s alpha generally increases as the intercorrelations among test items also increase. Because
intercorrelations among test items are highest when all items measure the same construct, Cronbach’s alpha is widely believed to indirectly indicate the degree to which a set of items measures a single construct.

**External Reliability**

External reliability is defined by how well the instrument is reproducible when repeated, either by the same observer (intra-rater reliability) or between observers (inter-rater reliability). Intra-rater reliability is generally tested by stability over time in a test–retest situation. For example, if one radiologist looks at the same set of mammograms on two occasions, how likely is that he/she will agree with himself/herself on his/her interpretation? Intra-rater reliability is demonstrated by the consistency of scores across different raters at a point in time. An example of this is when a number of radiologists were asked to interpret the same MRI scans in order to develop an MRI scoring system. Assuming that the items in this study were continuous variables (e.g., the 10-point progressive and the 20-point additive MRI scoring systems), the intraclass correlation coefficient (ICC) was the estimate used to test the inter- and intra-rater reliability.

**Intraclass Correlation Coefficient**

This is a measure of reliability of two or more continuous measurements made on the same subjects. It is a generalized measure of concordance adjusted for chance agreement between sets of measurements that are structured as groups, rather than data structured as paired observations. Often regarded within the framework of a repeated measures analysis of variance (ANOVA) or linear mixed effect models, it takes into account the three sources of random variation: subject, observer, and error.

\[
ICC = \frac{\text{Variance (Mean squares [MS] Subject)}}{\text{Variance (MS Subject)} + \text{Variance (MS Observer)} + \text{Variance (MS Error)}}
\]

**Case Study**

We will illustrate the ideas behind reliability of continuous data by examining a case study involving measurement of the superior limb of the hippocampus on MRI (Figs. 4.8, 4.9, 4.10, 4.11, and 4.12).

**Study Design**

MR images of 50 patients were obtained. They were presented in random order to one radiologist and in a different random order to a second radiologist. Each radiologist used the same imaging system and computer monitor to measure the length of superior, lateral, and inferior arms of the hippocampus on each side. They recorded their measurements to the nearest 0.1 mm.

**Results**

Across the 50 subjects, the readings done by the two radiologists differed by at most 0.2 mm. We can gain an informal notion of how reliable the measurements from the two readers were by comparing this to the range of measurements. The range of measurements for one reader was 0.7 mm to 2.6 mm and the range of measurements from the two readers was 0.7 mm to 2.6 mm and the range of

---

**Fig. 4.8** Measure of reliability of continuous data. The part of the MR image overlaid with calipers shows measurements at three locations on the left- and right-sided hippocampus. This image illustrates the difficulty with locating exact boundaries of the structures and shows why these measurements might be subject to measurement error.
measurements for the other reader was 1.3 mm to 3.4 mm. There was at least one image where the difference between the two measurements of the same structure (1.9 mm) was essentially as large as the difference in size across 50 people (1.9 mm for reader 1 and 2.1 mm for reader 2). Nevertheless, the difference in measurements of the same hippocampus is measurement error, whereas there may be a real variation in the size of the hippocampus whenever measurements are obtained by different readers. How can we formalize this weighing up of measurement error to true variation?

Understanding Intraclass Correlation Coefficient

To understand what it means, let’s examine the data that we collect when two readers each make a single measurement of the hippocampus on a series of MRIs, a study that will estimate inter-reader reliability. The observed measurement \( X_{ij} \) (the size of the hippocampus) made by a particular reader \( (j) \) on a single subject \( (i) \) can be decomposed into three parts:

\[
X_{ij} = \mu + \Delta_i + \varepsilon_{ij}
\]

where

- \( \mu \) = true average value for all subjects
- \( \Delta_i \) = true deviation of subject \( i \) from the average \( \mu \)
- \( \varepsilon_{ij} \) = measurement error on subject \( i \) read by reader \( j \)
Measurements: Validity, Reliability, and Responsiveness

In words, we can write

\[ \text{observed measurement} = (\text{true average}) + (\text{true deviation for subject}) + (\text{measurement error}) \]

Measurement error in this example includes both the between-reader variation (how far reader \( j \) is on average from the true value for subject \( i \)) and also within-reader variation (the effect of reader \( j \) making a single read). If reader \( j \) were to re-measure the hippocampus, there is no guarantee that the measurement would be the same. With only one measurement of each hippocampus by each reader, both of these sources of variation enter the measurement error \( \varepsilon_{ij} \).

**Fig. 4.9** illustrates an intraclass correlation coefficient (ICC) model of four readers considering that each reader assesses five subjects. If we had a perfect, error-free way of measuring subject one, for example, we would obtain the value \( \mu + \Delta_i \), the solid red dot. Differences between repeated measurements of the same value are due to measurement error; some measurements will be larger than this and some will be smaller. The actual values obtained from readers 1, 2, 3, and 4 on subject one are shown by the open red circles and can be shown algebraically as:

\[
\begin{align*}
X_{1,1} &= \mu + \Delta_i + \varepsilon_{1,1} \\
X_{1,2} &= \mu + \Delta_i + \varepsilon_{1,2} \\
X_{1,3} &= \mu + \Delta_i + \varepsilon_{1,3} \\
X_{1,4} &= \mu + \Delta_i + \varepsilon_{1,4}
\end{align*}
\]

where the terms \( \varepsilon_{1,1}, \varepsilon_{1,2}, \varepsilon_{1,3}, \) and \( \varepsilon_{1,4} \) can be positive or negative. To complete this model, we first assume that the deviations of individuals around the true mean have an average value of 0. This is another way of saying that the overall mean size across individuals is \( \mu \). We represent the between-person variance as \( \sigma^2_{\text{person}} \) and in the usual approach, we also assume that the between-subject deviations have a normal distribution, so that \( \Delta_i \sim \mathcal{N}(0, \sigma^2_{\text{person}}) \). Finally, we need to deal with the measurement errors \( \varepsilon_{ij} \). These differences, due to reader skill, judgment, and so on, are centered at zero and have a different variance, which we will call \( \sigma^2_{\text{error}} \). Assuming they are also normal, we have \( \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{error}}) \). A single measurement \( X_{ij} \) contains both a component of true between-person variability (how far
Before proceeding with the calculations, it is a good idea to plot the measurements from the two readers in a scatterplot (Fig. 4.10). The scatterplot in this figure shows that reader 2 has a tendency to have larger measurements than reader 1. In the model for the ICC that we presented above, this will result in a lower ICC than we would see if the points were just as scattered as they are below, but centered on the diagonal.

Another plot that is useful in presenting the variability between two measurements is called Bland-Altman plot (Fig. 4.11). In this graph, the x-axis reveals the average of the measurements taken by the two readers and the y-axis the difference in those two measurements.

Altman and Bland provided formulas for constructing upper and lower limits of agreement, an interval that should contain 95% of differences between two measurements. In the material for this chapter in the online supplement, R code is provided to make a Bland-Altman plot and calculate the limits of agreement. We use the function with the command below, where v1 and v2 let us supply the names to use for the two readers in the y-axis label.

\[
\text{ICC} = \frac{\text{Between Variance}}{\text{Total (Between + Within Variance)}} = \frac{s^2_{\text{person}}}{s^2_{\text{person}} + s^2_{\text{error}}}
\]

The true variance is compared to the total variance, which is composed of true variance and measurement error variance. The range of the ICC is between zero and one (0,1). It is near zero when the true variability is a small part of the total variability, a situation that indicates a relatively large amount of measurement error. It is near 1 when most variability is true variability and the measurement error is relatively small.

Because the y-axis and the x-axis are on the same scale, ranging over 0.25 units, Fig. 4.11 makes it clear that the differences between the two readers are just as large as the differences between measurements on different people. The Bland-Altman procedure suggests plotting a horizontal line at the average difference between readers; here we see that measurements for reader 1 are 0.04 units lower than measurements for reader 2. The limits of agreement of two measurements on the same person range from –0.136 to 0.056, further evidence of poor reliability when we notice that the average size is around 0.2.

We can use the ICC function in the R package psych to calculate the ICC if we arrange our data in a simple table. Each row contains one subject and each column represents a reader, as in the sample showing the first six subjects below:

<table>
<thead>
<tr>
<th>reader1</th>
<th>reader2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>0.21</td>
<td>0.13</td>
</tr>
</tbody>
</table>

If the data are stored in data frame or matrix hippo with just these two columns, then we type

\[
\text{> ICC(hippocampus)}
\]
to obtain the output below (some output has been omitted for brevity).

**Intraclass correlation coefficients**

<table>
<thead>
<tr>
<th>type</th>
<th>ICC</th>
<th>p</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single_raters_absolute</td>
<td>ICC1</td>
<td>0.044</td>
<td>0.378</td>
<td>-0.233</td>
</tr>
<tr>
<td>Single_random_raters</td>
<td>ICC2</td>
<td>0.203</td>
<td>0.017</td>
<td>-0.055</td>
</tr>
<tr>
<td>Single_fixed_raters</td>
<td>ICC3</td>
<td>0.297</td>
<td>0.017</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Number of subjects = 50 Number of Judges = 2

The ICC in the first row of the output is usually the one of most interest and is the one we described above. The ICC of 0.044 tells us that most of the variation is a result of discrepancies between the two readers. Only 4.4% of the observed variation is true person-to-person variation. The ICC is accompanied by a p value for a test of the hypothesis that the ICC is 0 (i.e., that all variation is random error), a hypothesis that in this case cannot be rejected. The final two columns are a 95% confidence interval (IC) for the ICC. Since the ICC cannot be less than zero, we would replace the lower bound by zero and report a 95% CI of 0 to 0.32. The second ICC (single random raters) removes the overall difference in mean scores between the two readers, so it does not include all the difference between readers in the estimate of measurement error. As a result, it is slightly higher than the first ICC. Finally, we can obtain an ICC that applies to the particular two readers in this study, rather than treating them as two possible readers from many that could have been selected. Again, this is less useful for most applications in radiology, where we would like to generalize the results from the two readers in the study to apply to any radiologists.

If we turn our attention to the measurements of the hippocampus on the left side of the brain, and repeat our analyses, we find that there is slightly better reliability. The scatterplot in Fig. 4.12a shows that points tend to lie closer to the diagonal and the Bland-Altman plot (Fig. 4.12b) bears this out; the mean difference is now closer to 0.025 and the limits of agreement run from –0.10 to 0.05. The output from the ICC function is shown below:

<table>
<thead>
<tr>
<th>type</th>
<th>ICC</th>
<th>p</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single_raters_absolute</td>
<td>ICC1</td>
<td>0.54</td>
<td>2.1e-05</td>
<td>0.31</td>
</tr>
<tr>
<td>Single_random_raters</td>
<td>ICC2</td>
<td>0.57</td>
<td>5.5e-08</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Our preferred ICC has a value of 0.54, with a 95% CI of 0.31 to 0.71. The interpretation of the ICC is that 54% of observed variability is person-to-person and 46% is measurement error. We can strongly reject (p = 0.00002) the hypothesis that the ICC is zero. The ICC that removes the average difference between readers is only slightly larger here, because the average difference between readers is smaller than for the right side.

The example above shows how to estimate reliability for two readers, the so-called interrater reliability. If the study design had involved one reader measuring each case twice—in a sense the second read taking the place of the second reader—we could have used identical methods for estimating reliability. In the latter design, we would have been measuring intra-reader reliability for this single reader. The only difference in the interpretation of the statistical model is that when we decompose the observed measurements this way:

\[ X_{ij} = \mu + \Delta_i + e_{ij} \]

the value \( e_{ij} \) represents the measurement error for the \( j \)th read on subject \( i \), when repeated reads are made by the same reader.
Research Methods in Radiology

When there are more than two readers or more than two reads per reader, it takes more space to present results in figures as they require plots of each pair of readers. For example, with four readers, there are six pairings. The ICC, however, does generalize to more than two readers or more than two measurement occasions. If we had a third reader, we would have a data set with three columns and run the ICC command exactly the same way.

> hippo
reader1   reader2   reader3
0.21      0.21      0.20
0.21      0.21      0.19
0.17      0.17      0.16
0.17      0.21      0.18
0.17      0.21      0.19
0.21      0.13      0.18
> ICC (hippo, missing = FALSE)

Finally, an experiment could involve multiple readers, each doing multiple reads. A common study design is that a set of images is read by two readers and then, at a later date, these same images are re-read in random different order by the same two readers. Such an experiment would allow us to simultaneously estimate inter-reader and intra-reader measurement error. While such designs can be efficient and give increased precision of estimates of both inter-reader and intra-reader ICCs, their analysis is beyond the scope of this book.

It is important not to confuse reliability, as measured by the ICC, with association, as measured by the Pearson or Spearman correlation coefficients. Two measurements can have a high correlation but exhibit poor reliability. For example, if one reader of the hippocampal sizes gave measurements that were exactly 1.5 cm larger than the other reader, the measurements would be perfectly correlated ($r = 1$) but would have zero reliability (ICC = 0).

### Kappa Statistic

The kappa coefficient is the test used to measure chance-corrected agreement for nominal and categorical variables. It measures the amount of agreement beyond what would be expected from chance alone. Like the ICC, the kappa coefficient is an estimator of chance corrected agreement. In small samples, observed agreement could even be worse than expected agreement, so kappa can be negative, but in most cases, kappa will range between 0 and 1. There are published guidelines on how to interpret kappa, but these simply translate a number to a word. In some areas of research (e.g., psychometry), 0.7 is considered to be a high kappa. In other fields, such as medical imaging, where money, time, and risk depend on a single diagnosis being correct, kappa = 0.7 may be considered unacceptably low. The kappa statistic is determined by the formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o$ is the observed amount (a proportion, not a percentage) of agreement among raters, and $P_e$ is the amount (probability) of agreement expected by chance, using observed data to calculate the probability that each observer randomly selects each category (Tables 4.1 and 4.2).

If there is perfect agreement, $P_o = 1$:

$$\kappa = \frac{1 - P_e}{1 - P_e} = 1$$

If there is no agreement in excess of chance, then $P_o = P_e$:

$$\kappa = \frac{P_o - P_e}{1 - P_e} = 0$$

### Table 4.1 Observed data from two radiologists making a diagnosis of liver nodules being present or absent for the scans of the same 51 patients

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>33</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>-</td>
<td>5</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>13</td>
<td>51</td>
</tr>
</tbody>
</table>
6.9 Measurements: Validity, Reliability, and Responsiveness

0.562. This is 0.66 or 66% of the potential improvement from chance agreement to perfect agreement; 66% of the possible improvement above chance has been attained.

The kappa statistic as described above also extends to multiple unordered categories. This is the statistic used when Doria et al. evaluated the sonographic patterns of the spleen in children of various ages. Let’s use this example, albeit with hypothetical data: two radiologists graded the ultrasound examinations of spleens using 1–3 scores (grade 1, granular pattern; grade 2, mild reticulonodular pattern; and grade 3, marked reticulonodular pattern), depending on their ability to identify hypoechoic nodules of various sizes (Fig. 4.13).

Let’s assume that they had rated a total of 125 spleens, and agreed that 25 were grade 1, 24 were grade 2, and 17 were grade 3 (Table 4.4).

Thus, the observed percentage agreement is:

\[ P_0 = \frac{25 + 24 + 17}{125} = 0.528 \]

We see what kappa means in this case. There is a potential 0.408 improvement from the observed agreement of 0.592 to perfect agreement (= 1). The observed agreement of 0.863 was 0.271 above the expected agreement of 0.592. This is 0.66 or 66% of the potential improvement from chance agreement to perfect agreement; 66% of the possible improvement above chance has been attained.

The kappa statistic as described above also extends to multiple unordered categories. This is the statistic used when Doria et al. evaluated the sonographic patterns of the spleen in children of various ages. Let’s use this example, albeit with hypothetical data: two radiologists graded the ultrasound examinations of spleens using 1–3 scores (grade 1, granular pattern; grade 2, mild reticulonodular pattern; and grade 3, marked reticulonodular pattern), depending on their ability to identify hypoechoic nodules of various sizes (Fig. 4.13). Please note that the investigators have used nominal/ordinal variables. The number of times they have scored each grade of spleen pattern was added up and entered into Table 4.3. Let’s assume that they had rated a total of 125 spleens, and agreed that 25 were grade 1, 24 were grade 2, and 17 were grade 3 (Table 4.4).

Thus, the observed percentage agreement is:

\[ P_0 = \frac{25 + 24 + 17}{125} = 0.528 \]
Table 4.3 Observed scores obtained from ultrasound examinations of children’s spleens

<table>
<thead>
<tr>
<th>Grade</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist 1</td>
<td>1</td>
<td>25</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>24</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>27</td>
<td>17</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>59</td>
<td>28</td>
<td>125</td>
</tr>
</tbody>
</table>

To calculate the probability of random agreement or concordance due to chance alone \( P_e \) the investigators had to calculate the number of expected chance agreements for each cell along the diagonal (highlighted cells) (Table 4.4). They had to multiply the corresponding row and column margin totals and divide by the total number of possible agreements, \( 125 \). These expected numbers were then added \( (12.2 + 16.0 + 11.4) \) to obtain the expected proportion of the agreement under the assumption the readers are just guessing, \( P_e = 39.6/125 = 0.317 \) (31.7%).

The kappa statistic is therefore:

\[
K = \frac{P_o - P_e}{1 - P_e} = \frac{0.528 - 0.317}{1 - 0.317} = \frac{0.211}{0.683} = 0.31
\]

The literature generally considers values < 0.21 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.\(^{23,24}\)

**Interpretation**

The results of this study showed that the inter-reader agreement was only fair, representing 31% of the possible gains above what would be observed if readers were guessing.

**Weighted Kappa**

Where categories are ordered (e.g., low, medium, and high grade cancer), there is a version of kappa that assigns partial agreement to categories that are “almost right.” This is called weighted kappa.\(^{25}\) For example, mistaking low grade for medium grade or vice versa might count as a partial agreement (e.g., 1/2 an agreement or some other value, depending on subjectively how far apart the categories are). Mistaking low grade for high grade or vice versa would count as a complete disagreement. The calculations for weighted kappa and for calculating confidence intervals for two-category, multicategory, and weighted kappa can be carried out using software (e.g., R package psych).

We can read in the data for Table 4.4, then use the `cohen.kappa` function to calculate both the unweighted kappa as previously described and the weighted kappa:

```r
> Spleen <- read.table("RadSpleen.txt",header=T)
> library(psych)
> cohen.kappa(Spleen)
```

Call: `cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha)`

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>estimate</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>unweighted kappa</td>
<td>0.19</td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td>weighted kappa</td>
<td>0.20</td>
<td>0.36</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Number of subjects = 125
measurements: validity, reliability, and responsiveness

4 Measurements: Validity, Reliability, and Responsiveness

validity, reliability, and responsiveness

degree to which an instrument measures what it is actually supposed to measure. validity can also be called accuracy. it is different from reliability as illustrated in figs. 4.14 to 4.21. in contrast to reliability which is affected by random error, validity is a function of systemic error or bias (fig. 4.7), and is affected by the same three main sources of systemic error.

there are several strategies that can be used to increase the validity of a study. if at all possible, making unobtrusive observations or using secondary outcome measures should eliminate subject bias. for example, counting the number of vials (total: 10 ml per vial) used for injection of a new MRI contrast agent (minimal dose: 30 ml) would be a more valid way to assess accuracy of contrast injection rather than directly asking subjects about the number of vials injected. calibrating instruments on a regular basis can be helpful if appropriate. finally, blinding the observers is an important strategy to reduce the effect of preconceived ideas that may influence their interpretation of the outcome.10

tests of validity are particularly challenging in the absence of a reference standard, and for many HRQoL constructs no true reference standard exists. however, the literature contains many already validated instruments that at least attempt to capture outcomes that may be of interest. using a combination of what already exists to validate an instrument or to demonstrate validity for your given population is often possible. different methods have been described in the literature to evaluate measurements when there is no reference standard. they include (1) imputing or adjusting for missing data on reference standard; (2) correcting available imperfect

validity

another way of looking at how well a measurement tool represents the truth is to assess the

using the software, we obtained the same value for kappa that was computed by hand but also received information on the 95% CIs. the upper limit of the CI is around 0.43, suggesting that at best the inter-reader reliability was fair to moderate. by default, the weighted kappa down weights agreement according to the square of how far apart the two values are but it is possible to use customized weights. here, we see that the weighted kappa of 0.36 (95% CI, 0.20–0.53) is slightly larger than the unweighted kappa. this results from partial credit being given to, for example, the values of 6 and 8 where the two readers were one category apart. the weighted kappa value for 3 x 3 (or larger) tables does not have the simple interpretation of the unweighted kappa as the fraction of above-chance agreement. although there are kappa formulas for more than two readers, the investigator does not have to present results for all pairs of readers, but can rather report a single measure of above-chance agreement among them. it is suggested that the investigator examines agreement between specific pairs of readers as subgroup analyses, for example, between two senior radiologists or between two trainees. finally, we should point out that kappa has been criticized for ignoring complexity in the data (e.g., for ignoring which levels readers agree on). to mitigate this limitation of kappa statistics, a large body of research, beyond the scope of this chapter, suggests the use of log linear regression models for modelling agreement.26,27

validity

another way of looking at how well a measurement tool represents the truth is to assess the

degree to which an instrument measures what it is actually supposed to measure. validity can also be called accuracy. it is different from reliability as illustrated in figs. 4.14 to 4.21. in contrast to reliability which is affected by random error, validity is a function of systemic error or bias (fig. 4.7), and is affected by the same three main sources of systemic error.

there are several strategies that can be used to increase the validity of a study. if at all possible, making unobtrusive observations or using secondary outcome measures should eliminate subject bias. for example, counting the number of vials (total: 10 μL per vial) used for injection of a new MRI contrast agent (minimal dose: 30 μL) would be a more valid way to assess accuracy of contrast injection rather than directly asking subjects about the number of vials injected. calibrating instruments on a regular basis can be helpful if appropriate. finally, blinding the observers is an important strategy to reduce the effect of preconceived ideas that may influence their interpretation of the outcome.10

tests of validity are particularly challenging in the absence of a reference standard, and for many HRQoL constructs no true reference standard exists. however, the literature contains many already validated instruments that at least attempt to capture outcomes that may be of interest. using a combination of what already exists to validate an instrument or to demonstrate validity for your given population is often possible. different methods have been described in the literature to evaluate measurements when there is no reference standard. they include (1) imputing or adjusting for missing data on reference standard; (2) correcting available imperfect

using the software, we obtained the same value for kappa that was computed by hand but also received information on the 95% CIs. the upper limit of the CI is around 0.43, suggesting that at best the inter-reader reliability was fair to moderate. by default, the weighted kappa down weights agreement according to the square of how far apart the two values are but it is possible to use customized weights. here, we see that the weighted kappa of 0.36 (95% CI, 0.20–0.53) is slightly larger than the unweighted kappa. this results from partial credit being given to, for example, the values of 6 and 8 where the two readers were one category apart. the weighted kappa value for 3 x 3 (or larger) tables does not have the simple interpretation of the unweighted kappa as the fraction of above-chance agreement. although there are kappa formulas for more than two readers, the investigator does not have to present results for all pairs of readers, but can rather report a single measure of above-chance agreement among them. it is suggested that the investigator examines agreement between specific pairs of readers as subgroup analyses, for example, between two senior radiologists or between two trainees. finally, we should point out that kappa has been criticized for ignoring complexity in the data (e.g., for ignoring which levels readers agree on). to mitigate this limitation of kappa statistics, a large body of research, beyond the scope of this chapter, suggests the use of log linear regression models for modelling agreement.26,27

validity

another way of looking at how well a measurement tool represents the truth is to assess the

degree to which an instrument measures what it is actually supposed to measure. validity can also be called accuracy. it is different from reliability as illustrated in figs. 4.14 to 4.21. in contrast to reliability which is affected by random error, validity is a function of systemic error or bias (fig. 4.7), and is affected by the same three main sources of systemic error.

there are several strategies that can be used to increase the validity of a study. if at all possible, making unobtrusive observations or using secondary outcome measures should eliminate subject bias. for example, counting the number of vials (total: 10 μL per vial) used for injection of a new MRI contrast agent (minimal dose: 30 μL) would be a more valid way to assess accuracy of contrast injection rather than directly asking subjects about the number of vials injected. calibrating instruments on a regular basis can be helpful if appropriate. finally, blinding the observers is an important strategy to reduce the effect of preconceived ideas that may influence their interpretation of the outcome.10

tests of validity are particularly challenging in the absence of a reference standard, and for many HRQoL constructs no true reference standard exists. however, the literature contains many already validated instruments that at least attempt to capture outcomes that may be of interest. using a combination of what already exists to validate an instrument or to demonstrate validity for your given population is often possible. different methods have been described in the literature to evaluate measurements when there is no reference standard. they include (1) imputing or adjusting for missing data on reference standard; (2) correcting available imperfect

using the software, we obtained the same value for kappa that was computed by hand but also received information on the 95% CIs. the upper limit of the CI is around 0.43, suggesting that at best the inter-reader reliability was fair to moderate. by default, the weighted kappa down weights agreement according to the square of how far apart the two values are but it is possible to use customized weights. here, we see that the weighted kappa of 0.36 (95% CI, 0.20–0.53) is slightly larger than the unweighted kappa. this results from partial credit being given to, for example, the values of 6 and 8 where the two readers were one category apart. the weighted kappa value for 3 x 3 (or larger) tables does not have the simple interpretation of the unweighted kappa as the fraction of above-chance agreement. although there are kappa formulas for more than two readers, the investigator does not have to present results for all pairs of readers, but can rather report a single measure of above-chance agreement among them. it is suggested that the investigator examines agreement between specific pairs of readers as subgroup analyses, for example, between two senior radiologists or between two trainees. finally, we should point out that kappa has been criticized for ignoring complexity in the data (e.g., for ignoring which levels readers agree on). to mitigate this limitation of kappa statistics, a large body of research, beyond the scope of this chapter, suggests the use of log linear regression models for modelling agreement.26,27

validity

another way of looking at how well a measurement tool represents the truth is to assess the

degree to which an instrument measures what it is actually supposed to measure. validity can also be called accuracy. it is different from reliability as illustrated in figs. 4.14 to 4.21. in contrast to reliability which is affected by random error, validity is a function of systemic error or bias (fig. 4.7), and is affected by the same three main sources of systemic error.

there are several strategies that can be used to increase the validity of a study. if at all possible, making unobtrusive observations or using secondary outcome measures should eliminate subject bias. for example, counting the number of vials (total: 10 μL per vial) used for injection of a new MRI contrast agent (minimal dose: 30 μL) would be a more valid way to assess accuracy of contrast injection rather than directly asking subjects about the number of vials injected. calibrating instruments on a regular basis can be helpful if appropriate. finally, blinding the observers is an important strategy to reduce the effect of preconceived ideas that may influence their interpretation of the outcome.10

tests of validity are particularly challenging in the absence of a reference standard, and for many HRQoL constructs no true reference standard exists. however, the literature contains many already validated instruments that at least attempt to capture outcomes that may be of interest. using a combination of what already exists to validate an instrument or to demonstrate validity for your given population is often possible. different methods have been described in the literature to evaluate measurements when there is no reference standard. they include (1) imputing or adjusting for missing data on reference standard; (2) correcting available imperfect

using the software, we obtained the same value for kappa that was computed by hand but also received information on the 95% CIs. the upper limit of the CI is around 0.43, suggesting that at best the inter-reader reliability was fair to moderate. by default, the weighted kappa down weights agreement according to the square of how far apart the two values are but it is possible to use customized weights. here, we see that the weighted kappa of 0.36 (95% CI, 0.20–0.53) is slightly larger than the unweighted kappa. this results from partial credit being given to, for example, the values of 6 and 8 where the two readers were one category apart. the weighted kappa value for 3 x 3 (or larger) tables does not have the simple interpretation of the unweighted kappa as the fraction of above-chance agreement. although there are kappa formulas for more than two readers, the investigator does not have to present results for all pairs of readers, but can rather report a single measure of above-chance agreement among them. it is suggested that the investigator examines agreement between specific pairs of readers as subgroup analyses, for example, between two senior radiologists or between two trainees. finally, we should point out that kappa has been criticized for ignoring complexity in the data (e.g., for ignoring which levels readers agree on). to mitigate this limitation of kappa statistics, a large body of research, beyond the scope of this chapter, suggests the use of log linear regression models for modelling agreement.26,27

validity

another way of looking at how well a measurement tool represents the truth is to assess the
reference standard; (3) combining multiple test results to construct a reference standard outcome including deterministic predefined rules, consensus procedures, and statistical modelling (latent class analysis); and (4) validating index test results.  

There are different types of validity, the most important being face, content, criterion, and construct validity.

**Face Validity**

Face validity or sensibility asks the common sense question, “Does the instrument make sense for a given population?” and is a subjective judgment that can be considered a form of content validity. Face validity is a concept that describes the practical value of the questionnaire with regard to length, scoring, content, and face validity. Questionnaires must have a purpose and framework consistent with the population to be studied. They must be comprehensive yet simple. The questions and instructions must be clear and unbiased, and allow for a range of responses. The output scale must be able to discriminate between meaningful changes in health states for the study population. The items must be relevant to the study population. Respondent and administrative burden must be minimized. For example, there is no point in using a 3-page 36-item questionnaire repeated every 3 days in a population of very ill patients. The patients will simply not use the instrument. This must also be considered when translating a tool, or using the same tool in another culture with a similar language, as different phrases may be interpreted differently.

**Conceptual Example**

Investigators plan to determine the best instrument to test the ability of adolescents to drive a car. They decide to investigate three potential instruments (**Fig. 4.14**).

**Example in Radiology**

Investigators plan to determine the diagnostic test (ultrasound biomicroscopy vs. gross specimen) that most closely resembles a histologic section of cartilage (reference standard) (**Fig. 4.15**).

**Content Validity**

Content validity concerns judgment about whether all important components of the

---

**Fig. 4.14** Face validity. Out of the three options of instruments to be used for testing the ability of adolescents to drive a car, the “on road test” is the one that makes more sense for this usage and, therefore, has greater face validity in this context.
Measurements: Validity, Reliability, and Responsiveness

Criterion Validity

Criterion validity is the degree to which the new measure compares to a reference standard, if one exists. There are two types of criterion validity: concurrent and predictive. Concurrent validity is the ability of an instrument (test measure) to correlate with a reference standard measure (typically used in clinical practice). Predictive validity is the ability of an instrument (test measure) to predict an outcome, even if this outcome is not the purpose of the scale. The new instrument's performance should be compared to existing validated measures, even if they do not represent a perfect reference standard, and differences discussed.
Example of Predictive Criterion Validity in Radiology

Investigators evaluated whether one or more of seven clinical variables were predictive of inaccurate ultrasound scanning (outcome) (Fig. 4.17).33

Conceptual Example of Predictive Criterion Validity

The Medical College Admission Test (MCAT) exam would have higher criterion validity if high MCAT scores predicted better grades in medical school.

Predictors of inaccurate ultrasound results in patients with clinically suspected appendicitis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Accurate result</th>
<th>Inaccurate result</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>12.8 ± 5.6</td>
<td>11.9 ± 2.1</td>
<td>&gt; 0.05 to 0.85</td>
</tr>
<tr>
<td>Females (%)</td>
<td>65 (38%)</td>
<td>106 (62%)</td>
<td>&gt; 0.05 to 0.15</td>
</tr>
<tr>
<td>Duration of pain (hours)</td>
<td>37 ± 25</td>
<td>42 ± 11</td>
<td>&gt; 0.05 to 0.16</td>
</tr>
<tr>
<td>Faces pain score (1–10)</td>
<td>2.1 ± 0.9</td>
<td>8.3 ± 1.1</td>
<td>0.01</td>
</tr>
<tr>
<td>Clinical probability of appendicitis (%)</td>
<td>81</td>
<td>70</td>
<td>&gt; 0.05 to 0.27</td>
</tr>
<tr>
<td>Ultrasound acquisition (trainee vs sonographer)</td>
<td>74</td>
<td>38</td>
<td>0.03</td>
</tr>
<tr>
<td>Ultrasound interpretation (trainee vs staff)</td>
<td>58</td>
<td>34</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Fig. 4.17 Criterion (predictive) validity. Let’s hypothesize a study whereby the investigators determined predictors of diagnostically inaccurate ultrasound scanning for suspected appendicitis. The main outcome measure was association between inaccurate ultrasound scanning (outcome) and age, sex, pain duration, Faces Pain Score-Revised, clinical probability of appendicitis, and ultrasound scanning, and ultrasound interpretation. In this case, inaccurate ultrasound scanning confirmed by surgery was considered the reference standard and the clinical variables that were compared with the reference standard were the predictors. The probability of inaccurate results for each predictor was calculated by using a fitted logistic model. (Modified from Schuh S, Man C, Cheng A, Murphy A, Mohanta A, Moineddin R, Tomlinson G, Langer JC, Doria AS. Predictors of non-diagnostic ultrasound scanning in children with suspected appendicitis. J Pediatr 2011;158:112–118, and now including hypothetical data.)
Conceptual Example of Concurrent Criterion Validity

Investigators developed a new psychological test designed to measure depression. The Green Depression Inventory is known to be a valid measure of depression. The investigators recruited a sample of individuals to take both the new Depression Scale and the Green Depression Inventory on the same day. Upon analysis of the results, the investigators find out that the higher the individual scores on the new Depression Scale, the higher their scores on the Green Depression Inventory (reference standard). Likewise, the lower the scores on the new Depression Scale, the lower the scores on the Green Depression Inventory. The investigators conclude that the scores on the new Depression Scale correspond to the scores on the Green Depression Inventory, thus establishing concurrent validity between the two scales.

Example of Concurrent Criterion Validity in Radiology

Investigators used pathological specimens of spleens as a reference standard measure to test the criterion validity of the distance between lymphoid follicles in the splenic parenchyma on ultrasound (test measure). The investigators assessed correlations between the test and the reference standard measures (Fig. 4.18).
In another example (Fig. 4.19), investigators compared T2 maps of articular cartilage of knees of a rabbit model of blood-induced arthritis (test measure) graded as “organized” or “disorganized” with grades of collagen organization assessed at corresponding histologic slides using polarized light microscopy (reference standard).

Construct Validity

Construct validity refers to how well a measurement conforms to a theoretical construct and can be measured by comparing scores to other measures that conceivably reflect the same theoretical construct. For example, if investigators developed a symptom assessment score, they could use “poor performance” on a performance status scale as a surrogate marker for increasing burden of disease based on a priori defined hypothesis. There are two types of construct validity: convergent and discriminant (or divergent). Convergent validity relates to how well two measures believed to reflect the same underlying phenomenon correlate or yield similar results. If two scales are valid methods for measuring a given outcome concerning convergent validity they should produce correlated scores. Discriminant or divergent validity relates to how well two measures believed to reflect different characteristics of an underlying phenomenon correlate or yield different results. If two scales are valid methods for measuring a given outcome concerning discriminant validity they should demonstrate low correlations.

Conceptual Example of Convergent Construct Validity

The results of an intelligence test are expected to correlate with the results of a test of neuronal activation.
Responsiveness

Once a measurement tool is developed and demonstrates to be valid and reliable, the final clinimetric property that must be shown before using a measurement as a research outcome is how sensitive the tool is to show change over time. It could be considered a type of validity but is an important enough property to stand on its own. Presumably one developed the tool to detect change in a research protocol, usually as a result of an intervention, and therefore this person would be interested in learning how the instrument would perform as a surrogate of response to change in order to determine what would be a clinically meaningful change.

A measurement is responsive if the value of the measurement changes when there are known changes in what it purports to measure. For example, suppose we measure lung tumor volume using MRI before and 6 months after a patient undergoes a course of radiotherapy for the tumor. Radiotherapy is expected on the average to shrink tumor volume. The MRI measurement of tumor volume is said to be responsive if it on average decreases by 0.41 to 0.60 (moderate correlation). Therefore, both the additive and progressive MRI scales proved to be valid measures for the proposed purpose for most of the clinical imaging constructs evaluated.

### Example of Convergent Construct Validity in Radiology

Investigators tested how well the MRI scores of severity of hemophilic arthropathy in elbow joints agreed with other markers of disease severity (Fig. 4.20).³⁴

### Conceptual Example of Discriminant Construct Validity

The results of an intelligence test are not expected to correlate with the results of a test of motor skills.

### Example of Discriminant Construct Validity in Radiology

Investigators tested whether physical examination, ultrasound, and radiographic scores were good discriminators of presence vs. absence of arthropathy in hemophilic knees, ankles, and elbows. They used areas under the curve (AUC) of scores to determine their accuracy to discriminate presence vs. absence of joint disease (Fig. 4.21).³⁴,³⁵

### Correlations between clinical and MRI parameters in patients with hemophilia

<table>
<thead>
<tr>
<th></th>
<th>Additive MRI Scale</th>
<th>Progressive MRI Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r (P value)</td>
<td>r (P value)</td>
</tr>
<tr>
<td>Severity of disease</td>
<td>0.41 (0.07)</td>
<td>0.32 (0.10)</td>
</tr>
<tr>
<td>Age</td>
<td>0.36 (0.04)</td>
<td>0.61 (0.03)</td>
</tr>
<tr>
<td>Number of prior joint bleeds in study joint</td>
<td>0.91 (0.01)</td>
<td>0.79 (0.03)</td>
</tr>
<tr>
<td>Pettersson X-ray score</td>
<td>0.76 (0.009)</td>
<td>0.71 (0.002)</td>
</tr>
<tr>
<td>Quality of life scores</td>
<td>0.81 (0.04)</td>
<td>0.65 (0.03)</td>
</tr>
<tr>
<td>Hemophilia Joint Health score</td>
<td>0.51 (0.10)</td>
<td>0.34 (0.50)</td>
</tr>
</tbody>
</table>

Fig. 4.20 Convergent (construct) validity. Clinical and imaging constructs (severity of disease, patient’s age, number of prior joint bleeds, Pettersson X-ray score, quality of life scores, and Hemophilia Joint Health scores) are compared with MRI constructs (progressive and additive scores) for testing the validity of the MRI scores in hemophilic elbows. The investigators hypothesized that if the proposed progressive and additive MRI scores were valid measures of joint status in hemophilia (in terms of construct validity), they would correlate with the clinical imaging constructs at least at the level of rho = 0.41–0.60 (moderate correlation). Therefore, both the additive and progressive MRI scales proved to be valid measures for the proposed purpose for most of the clinical imaging constructs evaluated. (Modified [using hypothetical data] from Doria AS, Lundin B, Miller S, Kilcoyne R, Dunn A, Thomas S, et al. Expert Imaging Working Group of the International Prophylaxis Study Group. Reliability and construct validity of the compatible MRI scoring system for evaluation of elbows in haemophilic children. Haemophilia 2008;14:303–314.)
Discriminant (construct) validity. Investigators hypothesized that physical examination Hemophilia Joint Health Scores (HJHS) and ultrasound (US) scores would be better discriminators of presence vs. absence of arthropathy in knees of patients with hemophilia than Pettersson radiographic scores. The areas under the curve (AUC) of physical examination (HJHS), ultrasound, and Pettersson X-ray scale represented the accuracy of these scores to discriminate presence from absence of arthropathy. The results show that ultrasound and HJHS scores are more accurate than X-ray scores for discrimination between presence and absence of arthropathy.

Fig. 4.21  Discriminant (construct) validity. Investigators hypothesized that physical examination Hemophilia Joint Health Scores (HJHS) and ultrasound (US) scores would be better discriminators of presence vs. absence of arthropathy in knees of patients with hemophilia than Pettersson radiographic scores. The areas under the curve (AUC) of physical examination (HJHS), ultrasound, and Pettersson X-ray scale represented the accuracy of these scores to discriminate presence from absence of arthropathy. The results show that ultrasound and HJHS scores are more accurate than X-ray scores for discrimination between presence and absence of arthropathy.

Conceptual Example

One widely used method of assessing internal responsiveness is to evaluate the change in a measure within the context of a randomized clinical trial involving a treatment that has previously been shown to be efficacious.

Examples in Radiology

In the aforementioned example on responsiveness of MRI measurements of the volume of lung tumors, the investigators may ask whether mean tumor size on MRI decreased after radiation in lung cancer patients, assuming that radiation on average decreases tumor size. Another example: Investigators examined the responsiveness of polarographic probes to measure short-term variations in oxygen partial pressure and blood flow in rabbit knees according to known changes in joint temperature (intervention). They obtained measurements from a time point prior to the start of the intervention, until completion of the experiment (when the tissue blood parameters stabilized a couple of minutes after the finalization of the temperature stimulus) (Fig. 4.22).

Many tests of internal responsiveness have been described, including the paired t-test, standardized effect size (SES, effect size I), standardized response mean (SRM, effect size II), and the responsiveness statistic (effect size III). The responsiveness statistic (the ratio of mean change after an intervention to the standard deviation of test scores of stable subjects) requires longitudinal data in stable subjects. The paired t-test statistic, the SRM, and the SES are commonly used measures of internal responsiveness, characterizing the ability of the instrument or questionnaire to measure change over time if change has occurred. Mean change scores becomes smaller over the 6-month time period. If the tumor volume measurements on MRI were on the whole unchanged, given what we know about radiotherapy and tumor volume, this would suggest that MRI is not picking up true changes and is not a responsive measurement.

Unfortunately, there is no consensus in the literature on how responsiveness should best be quantified. Two main types of responsiveness have been suggested: internal and external.

Internal Responsiveness

Internal responsiveness describes the ability of a measure to reflect a known change over a particular prespecified time frame. The research questions for this type of research design could be, “Do all the items within the scale change in the same direction, magnitude, and extent?” or, “Does the measurement tool measure the change in subjects where change is known to have really occurred?”
are calculated by subtracting the baseline scores from the follow-up scores for each patient and then calculating the mean. The paired t-test is used to test the hypothesis that there is no change in the average response on the measure at baseline compared to after the intervention. The t-statistic is the mean change score divided by its standard error (which is the standard deviation of the changes divided by the square root of the sample size).

\[
t = \frac{\text{Mean change score}}{(\text{standard deviation of change scores})/\sqrt{\text{sample size}}}
\]

With a reasonable sample size (n > 30), an absolute value of t > 1.96 allows the conclusion that a statistically significant (at p < 0.05) change in the measure has occurred. Effect size statistics were first proposed by Cohen\(^32\) and provide information on the
A measure that has a high level of variability in change scores in relation to mean change will have a small value for SRM. The SRM provides an estimate of change in the measure, standardized relative to the between-patient variability in change scores, that is, it indicates the size of the change in units of variation in within-person change. The SRM can be used, for example, to compare the responsiveness of scales, enabling the analysis of confidence intervals around the SRM, assuming that its distribution is approximately normal. In radiology, the SRM can be used, for example, to test the internal responsiveness of BOLD MRI to short-term changes in temperature of rabbit knees upon a controlled stimulus (increase and decrease of joint temperature by using a temperature selector; Fig. 4.20). In Nasui et al.39 the investigators hypothesized that BOLD MRI measurements would suffer moderate changes (SRM mean > 0.5 and ≤ 0.8) upon forced joint temperature changes.

The SES is less dependent on sample size than the SRM. If the baseline and follow-up standard deviations are equal, when the correlation between the baseline and follow-up scores is high, the SRM > SES; alternatively, when the correlation is low, the SES is higher than the SRM, and when the correlation is 0.5, the SES is equal to the SRM.

Note that neither of these measures of responsiveness produces a statistical test of the hypothesis that there has been no change. This would be accomplished by a paired t-test. Instead, the effect sizes quantify the amount of change in a measurement as a number of standard deviations. This allows the comparison of responsiveness of measurements that are on different scales. For example, if investigators repeat the aforementioned study using CT measurements of tumor volume and calculate the effect sizes, the measurement that has the larger effect size is said to be the more responsive.
4 Measurements: Validity, Reliability, and Responsiveness

External Responsiveness

External responsiveness reflects the extent to which changes in a measure over a specified time frame relate to corresponding changes in a reference measure. Thus, it requires that the investigators have an external measure of whether the underlying value changed which classifies subjects as “changed” or “unchanged.” Then investigators compare mean changes in measurements between subjects who have changed by the external standard and those who have not.

It answers the question: “Are changes in the instrument correlated with another measure that says the same thing?” In contrast to internal responsiveness, the measure is not in and of itself of primary interest. Rather, it is the relationship between change in the measure and change in the external standard. One motivation for this is that if the relationship is strong (i.e., the measure is shown to adequately capture changes in the standard), the measure may be used instead of the reference measure as an outcome in future clinical trials. Another motivation is that the measure under study should be a replacement for a standard measure, if the test measure is less invasive (e.g., non-radiation-bearing), less costly, and equally or more effective than the standard measure being used in clinical practice. The key feature of these external responsiveness measures is that a measure of true change exists and is assessed completely independently of the measurement of the change.

Because one investigates the external responsiveness of a measure by correlating the change scores with an external criterion and a reference standard for the magnitude of change is often unknown, this is often done by creating an external anchor.

Conceptual Example

If one investigator evaluates a symptom scale, for example, the patients can be asked to rate their change in symptoms after the intervention on a seven-item scale, where the bottom of the anchor would be “much worse,” the middle would be “no change,” and the highest point would be “most improved” (Fig. 4.23). In addition to providing an external measure of responsiveness, this approach would allow an exploration of the magnitude of change in scores in patients who had improved versus those who had not, providing a measure of “estimated change.”

Patient Opinion Sheet

Please circle which number best matches whether or not your symptoms have improved because of yesterday’s procedure:

1 2 3 4 5 6 7

Worst possible worsening of symptoms
No change in symptoms
Most possible improvement in symptoms

Fig. 4.23 External responsiveness. Example of an anchored question that can be used for the measurement of external responsiveness.
Examples in Radiology

Investigators examined the responsiveness of MRI to measure changes in cartilage scores of rabbit knees over a preestablished time frame (17 weeks) related to corresponding changes in a reference measure (gross specimen of cartilage at week 17) (Fig. 4.24).

Another example: Investigators compared changes in coronary artery flow on Doppler ultrasound in a clinical trial where treated and untreated groups are expected to present with different magnitudes of change in arterial flow Doppler values over time. The investigators compared the mean change...
in patients who were revascularized to the mean change in flow in patients who were not revascularized.

Many tests of external responsiveness have been described. One approach recognizes this situation as being analogous to a diagnostic test study. The investigators have two groups (truly changed and unchanged) and a continuous variable that acts similarly to a diagnostic test that classifies subjects to one of the two groups. In this view, the analysis is similar to the analysis of a continuous-value diagnostic test. The responsiveness of an instrument is summarized as the area under the receiver operating characteristic (ROC) curve formed by assessing possible threshold values in the change score. In the ROC method, responsiveness is described by its sensitivity and specificity in detecting improvement or deterioration. Because it uses a dichotomous outcome of improvement vs. nonimprovement, it does not provide precise information about the degree of deterioration or improvement.

Another approach relates to correlation that indicates how changes on two measures vary together, lying between −1 and +1, with positive values indicating a positive association and negative value, a negative association. The attractiveness of the correlation coefficient in assessing responsiveness is that it examines whether a measure is responsive relative to a specific alternative outcome. This reflects an interest in how well changes in one measure predict changes in another. Correlational analysis links naturally with these regression models but regression models have the potential to be more informative.

A third commonly used approach treats the change from baseline to follow-up as the outcome in a regression model with the known change groups as the predictor variables. The measure of responsiveness is the estimated difference in the change score between the “changed” and the “unchanged.” This difference can be divided by the between-subject baseline standard deviation, as in ES(I), so that it is measured in units of between-person variability and can be compared between measurements on different scales. Regression models generate an easily interpreted index in the form of the regression coefficient $b$. Furthermore, one can carry out a goodness-of-fit assessment to check the plausibility of the regression model. It is possible that the relationship between a new measure and a traditional clinical measure is not linear, and in this case the model may be generalized. Therefore, regression models have advantages over simple correlates for assessment of external responsiveness.

Finally, some validated instruments have a reported clinically important difference, to which the observed change scores could be compared. For example, for the Quality of Life (QLQ)-C30 questionnaire, which assesses the quality of life of patients with lung and breast cancer, a 10-point change in mean score on the 100-point scale is considered a small but meaningful clinical response, and a 20-point difference a moderate response.

The key feature of the external responsiveness measure is that a measure of true change exists, which is assessed completely independently of the measurement of the change. If investigators are assessing the external responsiveness of change in the example of vascular flow measured by Doppler sonography, then the formation of the “true” groups “changed” and “unchanged” cannot in any way include the Doppler sonography measurements as part of a composite assessment of whether a patient has changed or not.

### Highlights of Key Points

- Four main types of data are used for measurements of clinical phenomena: nominal, ordinal, interval, and continuous.
- Measurements are subject to three potential sources of error: those related to the subject (examinee), the examiner or tester, and the examination (instrument).
- Variability of measurements can be attributed to random error (affecting reliability), and some can be attributed to systematic errors such as bias or confounding variables (affecting validity).
• The following questions answer different concepts for validity, reliability, and responsiveness:
  a. Validity (overall concept): “Is the scale measuring what we think it is?”
  b. Face validity: “Does it look like it measures what it intends to measure?”
  c. Content validity: “Does it include representative samples of the construct we want to measure?”
  d. Construct validity (overall concept): “Does the instrument express how components of a measure relate to each other and to other measures according to a predetermined hypothesis (mini-theory)?”
  • Convergent validity: “Does our construct measure correlate with similar, related constructs?”
  • Discriminant validity: “Does our construct measure correlate with dissimilar, unrelated constructs?”
  e. Criterion validity: “Does our construct measure correlate or predict a directly observable phenomenon?”
  • Concurrent validity: “Does it correlate with a reference standard measure when both are administered at the same time?”
  • Predictive validity: “Does it predict an outcome of interest?”
  • Reliability: “Does the measure of a stable phenomenon get similar results by different people and instruments at different times and places?”

Continuous Variables
• If measurements can be paired (reader 1 vs. reader 2, reader 1 at two different times):
  • Make a scatterplot.
  • For each pairing, make a Bland-Altman plot of the difference vs. the average.
  • Calculate the ICC and its 95% CI.
  • Interpretation of ICC: It relates to the proportion of variability that is true variability, not reader error or measurement error. Values closer to 1 indicate better reliability. Values closer to 0 indicate worse reliability.
• Pearson and Spearman correlation coefficients do not measure reliability.

Categorical Variables
• Make a table of the findings for one reader versus the other (for inter-reader reliability) or for one reader (for intra-reader reliability) using the first read versus the second read.
• Calculate percent agreement. For binary variables, the kappa value indicates the above-chance agreement. For categorical variables with three or more levels, one can calculate unweighted kappa as a measure of above-chance agreement. If the variable is ordered, one can calculate an unweighted or a weighted kappa, which credits part-agreement.
• There are extensions of kappa to multi-reader settings.
• Responsiveness (= sensitivity to change): “Does the measure of a phenomenon change as conditions change?”

References


25. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968;70:213–220


Learning Objectives

- To understand what is meant by observational design.
- To understand the difference between retrospective and prospective research designs.
- To understand the various types of observational designs including cross-sectional, retrospective cohort, and case-control designs.
- To understand threats to the validity of observational studies including confounding and effect modifiers.
- To understand strategies for reporting observational designs.

Introduction

Nonexperimental Designs

This chapter focuses on observational or “nonexperimental” designs. Nonexperimental methods are approaches whereby the investigator does not alter or manipulate the circumstances of the participants in the research study. Nonexperimental designs can be comparative, such as case-control studies, or noncomparative, such as case series. However, distinct from experimental studies (randomized controlled trials [RCTs]), in observational designs randomization of subjects to treatment groups, interventions, or exposures does not occur. For example, an investigator may be interested in comparing two treatments or procedures and retrospectively reviews the outcomes of patients who received these interventions. While this study is comparative, the researcher has no impact on how decisions were made in terms of which patients received the treatments or procedures being compared, nor does the investigator have control over how they were administered.

Although nonexperimental designs have low internal validity and are generally not considered the optimal approach to compare therapies or procedures, they are frequently used. One of the reasons for this is that in some circumstances observational studies are the only practical strategy. This may occur when an RCT is impractical, for example, when the condition under investigation is rare. Observational approaches may also occur, earlier in the experience with a new treatment, where the ethical imperative of equipoise (evidence for similar effectiveness of both treatments to be compared) is not yet met for an RCT to be undertaken. Results from observational designs may in fact form the justification for RCTs. In addition, observational studies are less expensive than experimental studies and are therefore the most commonly utilized research design in health care. Observational studies may also be more reflective of “real life” as they often have less restrictive inclusion criteria than experimental designs and often include the entire spectrum of patients with a particular illness, rather that selecting patients with specific characteristics which limits the generalizability of the results. 1

Retrospective versus Prospective Studies

The distinction between retrospective versus prospective studies is often assumed to be inherent to a particular research design. While this is partly true, in that some types of studies can only be prospective, for example, a clinical trial, many study designs can be done in both a retrospective or prospective manner, for example, a cohort study.
A retrospective study is a study planned after the treatment (and often after the outcome has occurred). This is therefore a study whereby the investigator has no control over what variables will be available for collection, as well as how the treatment will be performed. The primary advantage of a retrospective design is that it is faster to perform, as all treatments and outcomes have occurred. A researcher only needs to collect and analyze the data. Retrospective studies are also often cheaper and faster than prospective studies. Retrospective studies are very useful in rare diseases, whereby prospective enrollment and data collection may be unfeasible.

In contrast to a retrospective study, a prospective study is one that is planned prior to the treatment (and outcome) occurring, allowing the investigators to have greater control over the treatments or exposures evaluated in the study as well as what data will be collected. Since the specific variables to be collected can be determined in advance, the investigator can ensure that they have the necessary data in order to best evaluate the research question that they are investigating. On the other hand, since none of the treatments (or outcomes) have occurred or data are available at the outset of the study, the researcher may need to wait a number of years before they are able to address the question that they wish to answer. Therefore, the completeness of the data afforded by a prospective study needs to be balanced against the additional time and cost of this approach.

Types of Nonexperimental Designs

There are different types of observational studies whose results yield different levels of evidence: case report and case series, cross-sectional, cohort, and case-control studies. Case reports and case series offer lower evidence than cross-sectional, cohort, and case-control studies. Prospective studies generally afford a higher level of evidence than retrospective designs. Evidence from nonexperimental designs is on a lower level than experimental designs, such as RCTs. Whenever possible, the strongest design should be used.

Although studies evaluating diagnostic tests are an important type of nonexperimental design in radiology, this chapter focuses on other types of nonexperimental designs. Diagnostic tests are covered in Chapter 3.

For each of the main types of nonexperimental designs, we briefly describe the key features of the design and provide an example of that type of design. While each of the designs mentioned is considered separately, studies may be a blend of different types of nonexperimental methods.

Case Reports and Case Series

The simplest observational study designs are case reports and case series. The key role of these studies is to report findings of a rare disease or a novel observation or the presentation of a common condition in one or more patients. The results of the studies generate hypotheses that require subsequent formal testing and are therefore considered to be “hypothesis-generating studies.” These studies are then used to design subsequent studies to confirm the observation seen in the case series. Without subsequent confirmatory studies, observations made in case series cannot be generalized to a larger population of patients. In radiology, case reports and case series can be educational tools displayed as pictorial reviews of the imaging characteristics of diseases or group of diseases, such as those published in Radiographics which represent an example of a case series.

Cross-Sectional Designs

Cross-sectional designs relate to observations of variables in a population where all measurements are obtained on a single point in time. Therefore, no follow-up period is expected. Subjects are not selected based on exposure or outcome since both occur at the same time point (Fig. 5.1). Studies that use this design are relatively inexpensive and easy to carry out since there is no waiting for outcomes to occur. They can only establish an association, not a cause–effect relationship, and are therefore not practical for assessment of rare diseases and cannot estimate disease incidence.2,3,4
Cohort Designs

In cohort designs, subjects are selected based on an exposure of interest (e.g., a type of intervention procedure, a disease, a drug, an imaging technique). Subjects are then assigned to a group or cohort (exposed vs. nonexposed) based on this. Both exposed and nonexposed cohorts are then followed backward (if retrospective) or forward in time (if prospective) to verify how many patients in each group develop/did develop the outcome of interest. The exposure is an independent variable and the outcome, a dependent variable. This design is suitable to explore causality and can be used to evaluate incidence, cause, and prognosis.

For example, if investigators are interested in examining whether a CT scan during childhood leads to future impairment in cognitive function, they could perform a cohort study. In this case, the investigators decide to perform a retrospective study whereby they enroll 4,000 15- to 18-year-old adolescents attending grades 10 to 12, collect data on the number of head CT scans that they had undergone in their childhood (0–7 years), and obtain tests of cognitive function on these children. The investigators had hypothesized that the radiation from previous CT scans could be associated with future reduced cognitive function and this hypothesis was tested using the data available.

The steps performed by the investigators as part of this study were as follows:

1. Identify a suitable cohort: Ask questions to the population of interest on previous medical history which can be systematically collected and recorded.
2. Collect data about predictor variables: Ask questions about the cohort’s previous medical history (prior CT scans) and other potential risk factors.
3. Collect data about subsequent outcomes that occurred at a later time: Retrospectively explore the relationship between cognitive performance (outcomes) and history of head of CT scan in childhood (exposure).

This same question could be addressed with a prospective cohort study, although the steps...
required for a prospective study would be slightly different, as follows:

1. Assemble a cohort: Enroll children of a certain age range (e.g., 0–7 years) who would have the potential to undergo CT heads (for head injury, seizures, etc.) in the future during the conduct of the study.

2. Measure exposure variable(s) and potential confounders: Follow-up on the number of CT scans and other potential risk factors that the cohort are exposed to during the follow-up period. This could be done by review of the medical records, or ongoing contact with the subjects (e.g., periodic telephone calls or surveys).

3. Follow-up to measure outcomes: Follow-up the patient cohort until the age of 18 years and collect data on cognitive performance at age 18.

As you can see, while the retrospective and prospective cohort studies address the same question, the prospective study would take many years to complete, whereas the retrospective study could be fairly quickly done. However, recall of number of head CT scans done in childhood may be difficult for parents and especially for adolescents themselves. Therefore, this may limit the accuracy of the data collected as part of the retrospective study. The prospective study may allow for more complete data, although one of the major challenges of a long-term study is loss to follow-up and attrition, whereby subjects who enroll in these studies may move or decide that they no longer want to be part of the study over time.
Observational Designs

Case-control studies are a design that aims to compare groups with and without an outcome of interest and to explore the role of different exposures and the potential association between these variables and outcome of interest (Fig. 5.3). Since cases (experienced outcome) and controls (did not experience outcome) are defined by the occurrence of the outcome, by definition case-control studies are all retrospective as the outcome has to have occurred prior to the study beginning and exposures all must be in the past. Case-control studies are a design, therefore, that can be used to identify associations between exposures and outcome in rare diseases, or in conditions with rare outcomes or when there is a long time lag between the exposure and the outcome.

Case-control studies are relatively inexpensive, but are vulnerable to unmeasured confounding variables. Confounding will be discussed later in this chapter and in Chapter 6. This can be addressed, to some extent, by matching cases and controls on particular characteristics. Matching refers to enrolling controls in the study who have similar characteristics as cases, such as age or sex. However, variables used to match cases and controls cannot be included in analyses as potential predictors as the investigator has determined the distribution of these variables within the groups. Similarly, since the number of subjects who experience an outcome is determined by the researcher (i.e., number of cases enrolled relative to number of controls enrolled), case-control studies are not suitable for estimation of incidence of disease. Case-control studies are, however, useful for exploring associations between exposures and outcomes and for hypothesis generation for further studies to explore the relationships seen in case-control studies.

As an example, an investigator is interested in exploring the relationship between cognitive impairment in Canadian children and cranial radiation during childhood. The decision to perform a retrospective vs. prospective study is partly dictated by the research question as well as available resources. In this particular example, since head CT during childhood is a relatively infrequent occurrence, one would likely need to enroll many children to allow for a sufficient number in the cohort who had exposure to cranial radiation to allow for sufficient power to detect a difference in outcomes between those who were and were not exposed. Therefore, such a study would be tremendously expensive to do in a prospective manner, and perhaps, despite the limitations of the retrospective approach, it may remain preferred for this particular research question.

**Fig. 5.3** (a) Time points for exposure and outcome in case-control studies. (b) The data collection regarding predictors (prior performance of CT scans for head injury in childhood) and outcomes (cognitive impairment) for cases and controls occurs after the outcome has occurred.
and exposure to head CT. These investigators have identified 332 subjects, between 15 and 18 years old, with cognitive impairment (intelligence quotient [IQ] < 85) who will be designated as cases. The investigator has identified 3,668 age- and sex-matched control subjects from an existing database without cognitive impairment (IQ > 100). The investigators collect data on both cases and controls on a number of variables that may be associated with cognitive deficits in adolescence, including whether the subject had undergone head CT scans between ages 0 and 7 years.

The steps performed by the investigator as part of this study were as follows:

1. Select cases: Identify 332 adolescents with cognitive deficit from institutional database.
2. Select controls: Identify 3,668 adolescents without cognitive deficit matched for age and sex.
3. Measure the predictor variable: Look back to determine how many individuals (cases and controls) were exposed to the potential predictor (head CT) in childhood.

Case-control studies can be performed on data collected in a cohort study, in what is referred to as a “nested case-control study” because the case-control study is nested within the cohort study. In a cohort study both exposed and nonexposed subjects are included and data collected on them as part of the follow-up for the cohort study. However, within the cohort relatively few will experience the disease (i.e., outcome). Using the data of those who did and did not experience the outcome, the investigator can define cases and controls and perform a nested case-control study. One of the advantages of this approach is that since cases and controls are part of a prospective cohort study, the data available may be more complete than had it just been collected as part of a retrospective case-control study rather than a nested case-control study. Nested case-control studies are potentially powerful designs and important findings can be derived from them. For example, the relationship between mammographic density and breast cancer risk was investigated as part of a nested case-control study using data from the 14,291 women in the New York University Women’s Health Study, a cohort design. Investigators estimated the relations of mammographic patterns/densities and breast size to breast cancer risk using the archived mammograms of 197 women who developed breast cancer (cases) and of 521 age- and sex-matched controls from same cohort. Using this design, a significantly increased risk for specific mammographic patterns was found.7

**Bias and Confounding**

Bias is an error in research methodology that may cause the results of the study to deviate from the “truth.” Although numerous types of bias have been reported in the literature,6 the main sources of bias in radiology studies are the patient, the intervention/exposure, the data gathering, and the interpretation/reporting approaches. Bias can occur at any point during the study including recruitment, intervention, data collection, analysis, and publication (Table 5.1).9,10,11,12,13 It is important to understand the potential sources of bias when designing a study, so that their impact can be minimized. While there are some advanced statistical techniques that can be used to ameliorate the impact of bias introduced by errors in study design, it is best to avoid introducing the bias as it may not be possible to minimize the impact of bias on the results once introduced.

**Confounding**

Confounding is a particular type of bias that results from a factor being introduced that is associated with the exposure and the outcome, but is not part of the causal pathway between the exposure of interest and the outcome, yet impacts the outcome (Fig. 5.4).14 The concept of confounding is discussed in Chapter 6. As discussed in that chapter, confounding is a particular concern in nonexperimental designs, whereby factors other than random chance determines which treatment a subject receives and it is these factors (confounders) that may impact outcome, rather than differences in the treatments received. This type
# Table 5.1 Bias in observational research in radiology

<table>
<thead>
<tr>
<th>Type</th>
<th>Systematic error/bias in</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus: Patient/exposure-intervention</strong></td>
<td></td>
</tr>
<tr>
<td>• Selection bias</td>
<td>Including study subjects (sicker, milder cases, health care workers, volunteers)</td>
</tr>
<tr>
<td>• Sampling bias, referral bias</td>
<td>Some members of society are more likely to be included, referred to the center than others (profession, socio-economic status, access)</td>
</tr>
<tr>
<td>• Image-based selection bias</td>
<td>Study enrollment mandated a specific image, patients are included based on the availability of such imaging study, study population is selected from a true target population</td>
</tr>
<tr>
<td>• Study examination bias</td>
<td>Study enrollment limited to technically excellent studies, resulting in overestimation of sensitivity and specificity (removal of false-negatives and false-positives, respectively)</td>
</tr>
<tr>
<td>• Disease spectrum bias</td>
<td>Within patient groups, one end of the spectrum gets investigated only</td>
</tr>
<tr>
<td>• Self-selection bias, “healthy volunteer bias”</td>
<td>Study enrollment on the basis of self-selection, limits generalizability</td>
</tr>
<tr>
<td>• Channeling bias</td>
<td>Patient prognostic factors or degree of illness leads to inclusion into one study over another</td>
</tr>
<tr>
<td>• Participation bias</td>
<td>Unequal response to additional factors required to actually join a study (distance to center, financial burden, other personal constraints)</td>
</tr>
<tr>
<td>• Transfer bias/loss to follow-up bias</td>
<td>Unequal loss to follow-up between groups gets treated similarly in the analysis</td>
</tr>
<tr>
<td>• Language bias</td>
<td>Inappropriate definition of the eligible population</td>
</tr>
<tr>
<td>• Confounding by indication</td>
<td>Association with the exposure without being the consequence of the exposure and with the outcome independently of the exposure</td>
</tr>
<tr>
<td><strong>Focus: Data generation/gathering</strong></td>
<td></td>
</tr>
<tr>
<td>• Recall bias</td>
<td>Outcomes of treatment rely on subjects’ recollections of events prior to or during the treatment process, information of exposure is systematically classified differently between groups</td>
</tr>
<tr>
<td>• Information bias</td>
<td>Wrong, incomplete, or inexact recording of variables</td>
</tr>
<tr>
<td>• Interviewer bias</td>
<td>Interviewer makes systematic distinctions on how information is solicited, recorded, or interpreted between patient groups, typically when the interviewer is also an investigator</td>
</tr>
<tr>
<td>• Misclassification bias</td>
<td>Exposure itself is poorly defined or if proxies of exposure are utilized, leading to wrong assignment to grouping</td>
</tr>
<tr>
<td>• Verification bias, workup bias</td>
<td>Systematic difference in the manner in which the disease status is defined between groups causes: dependency of tests, unblinding to results of reference standard</td>
</tr>
<tr>
<td>• Performance bias</td>
<td>Experience of one or more people delivering exposure measures systematically different from experience of remainder operators or readers</td>
</tr>
<tr>
<td>• Chronology bias</td>
<td>Historic controls are used as a comparison group</td>
</tr>
<tr>
<td>• Follow-up bias/medical surveillance bias</td>
<td>Groups have systematically different follow-up regimens based on test results</td>
</tr>
</tbody>
</table>

*(continued)*
Table 5.1 Bias in observational research in radiology (Continued)

<table>
<thead>
<tr>
<th>Type</th>
<th>Systematic error/bias in</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Response bias</td>
<td>Missing data are not randomly distributed, disease-free cases were less thoroughly investigated, motivation to respond to all questions may differ between groups</td>
</tr>
<tr>
<td>• Contamination bias</td>
<td>Execution of an intervention, by using one agent as part of a diagnostic test (e.g., contrast material), may impact the result of another diagnostic test</td>
</tr>
<tr>
<td><strong>Focus: Interpretation/reporting</strong></td>
<td></td>
</tr>
<tr>
<td>• Reviewer bias</td>
<td>Inappropriate blinding of person reviewing the results, reviewer of a new test may be aware of the results of the reference test</td>
</tr>
<tr>
<td>• Diagnostic-review bias</td>
<td>Reference tests are not definite, reviewer utilizes the new diagnostic test result to modify the interpretation of the reference test</td>
</tr>
<tr>
<td>• Test-review bias</td>
<td>Diagnosis is known to reviewer at time of review, which may influence the test interpretation</td>
</tr>
<tr>
<td>• Incorporation bias</td>
<td>Reviewer incorporates the new diagnostic test result to make a diagnosis</td>
</tr>
<tr>
<td>• Imperfect standard bias</td>
<td>Reference tests are imperfect, therefore some study subjects will have a second, surrogate reference test based on their characteristics (socioeconomic status, age)</td>
</tr>
<tr>
<td>• Reader-order bias</td>
<td>Reviewer retained knowledge of the first test and incorporated it in the interpretation of the second test</td>
</tr>
<tr>
<td>• Measurement bias</td>
<td>Discrepancy in the measurement approach between two groups of tests</td>
</tr>
<tr>
<td>• Cluster bias, repeat measurement bias</td>
<td>Multiple measurements are taken from the same subject and are considered independent</td>
</tr>
<tr>
<td>• Context bias</td>
<td>Altered disease prevalence influences test characteristics, important for interpretation of equivocal test results</td>
</tr>
<tr>
<td>• Publication/citation bias</td>
<td>Researchers and sponsors unwilling to publish unfavorable results</td>
</tr>
</tbody>
</table>

of confounding is referred to as confounding by indication, and is a key reason why randomization, by avoiding confounding by indication, allows for unbiased assessments of treatment response. Although randomization is one method to control for confounding, there are other approaches that can be taken such as specification (restricting enrollment of subjects with only certain characteristics), matching of subjects on potential confounders as discussed in case control studies, and various analytic techniques.

The issue of confounding by indication is of concern in the previous example provided for the case-control study. In this study investigators enrolled adolescents with cognitive impairment (cases) as well as controls without cognitive impairment and retrospectively explored the association between cognitive impairment in adolescence and performance of head CT scans in childhood. In this hypothetical example, investigators determined that those with cognitive impairment were more likely to have had a head CT as a child. It is important to recognize that since this is a case-control study, this finding is merely an association and one should not conclude that the performance of head CT scans leads to cognitive impairment. In fact, it is likely that this result can be explained by confounding by indication, whereby adolescents with cognitive problems were more likely to have undergone a head CT scan and that these same adolescents are more
5 Observational Designs

Confounding

Exposure — Outcome
Third variable (confounder)

Condition #1 Condition #2

Confounding Example

Brain radiation (radiotherapy) in childhood — Cognitive deficit in adolescence

Brain surgery in childhood

Third variable

Fig. 5.4 (a) Illustrates the two conditions for a third variable or extrinsic factor to be a confounder. Condition 1, the third variable (potential confounder) is associated with the exposure without being the consequence of the exposure; and condition 2, the third variable (potential confounder) is associated with the outcome independently of the exposure (it is not an intermediary). (b) Brain surgery in childhood (confounder) is correlated with brain radiation (exposure) in patients with cancer and is a risk factor for cognitive deficit in adolescence (outcome)—even if the amount of brain radiation received in childhood was low—therefore it holds all conditions to be considered a confounder.

How to Control for Confounders

Conditioning is the epidemiologic term used to define strategies available to control for confounding (see Chapter 6) which include stratification, specification (or sample restriction), randomization (see Chapter 6) and matching in the study design phase, and adjustment (model fitting) in the study analysis phase. Details on model fitting are available in Chapter 14. Table 5.2 summarizes the methods for controlling for selection bias and confounding in observational (and experimental) studies.

In the design phase:

Specification and matching: These are strategies that involve changes in the sampling scheme. They prevent confounding by allowing comparison of cases and controls that share comparable levels of the confounder.

Specification or sample restriction refers to a strategy in which eligibility for entry into a study is restricted to individuals who possess a narrow range of characteristics. This strategy simply excludes every subject who holds a characteristic (variable) distinct from the ones a priori specified. Example: Subject’s age: If younger children are expected to have thicker articular cartilage than older children and cartilage thickness affects cartilage anisotropy on diffusion-weighted MRI (outcome), then the investigators should decide on accruing subject within a specific age range (15–18 years) where no significant differences in cartilage thickness are expected. Using the restriction approach, only patients within this age group should be included in the study. This approach makes the study sample more homogeneous, but limits its generalizability, in that the results are not applicable to subjects with a different range of characteristics.

Anticipated problems include:

- There may be effect modification (e.g., the more exercise subjects practice in daily life, the more anisotropy is expected in the cartilage).
- Restriction limits generalization (clinicians may need information about anisotropy of
Anticipated problems include:

- Investigators cannot look at the matching variable as a predictor.
- There is potential for “overmatching.”
- Challenging to match on many variables.

“Overmatching” in case-control studies can be a design fault and may lead to bias. If investigators match on variables that affect or are affected by the exposure under study (but are not associated with the outcome), this may result in an increased chance that the matched case and control have the same history of exposure; therefore, the two variables are essentially measuring the same thing or the second variable is a consequence of the first one. Overmatching can bias the study toward finding no effect, thus diminishing the ability of a study to detect an increased or decreased estimated relative risk.

cartilage for different pediatric age groups, not only for 15- to 18-year-old patients).

Matching is a hallmark of case-control studies. If some subjects’ characteristics seem strongly related to either exposure or outcome, and these confounding factors cannot be controlled by randomization, investigators should be sure that they are comparable in cases and controls, which can be achieved by the “matching” strategy. Although it is usually applied in pairs of cases and controls (pairwise matching), it can also be done in groups (frequency matching). Example: If age is a confounder in a study and adolescents with the outcome (cases) have ages ranging from 15 and 18 years then matching by confounder (age-matching) is required. As a result, adolescents without the outcome (controls) in this study should also be 15 to 18 years old.
Observational Designs

Example


This was a case-control study of workers at a nuclear reprocessing plant in which over-matching obscured the relationship between cumulative radiation dose and mortality from leukemia. We will review the key aspects of the study design in order to illustrate the concepts on over-matching.

Study Design: In this study the disease of interest was leukemia, and the risk factor in question was cumulative occupational external radiation dose.

Study Sample and Objectives: The authors examined the relationship between dose accumulated by workers from 1976 to 1983 and specific mortality and investigated the hypothesis that dose was positively associated with risk.

Matching Strategy: According to the authors, matching on sex and date of birth was necessary because of the underlying difference of the risks of leukemia between the sexes, and with changing age. Matching on date of entry was thought necessary because the risk of leukemia changed with calendar time, and this was intended to eliminate this factor. However, dose also changed with calendar time.

Results and Interpretation: The results of the fully matched datasets (current study) contradicted those of previous studies that found a significant positive association between mortality from leukemia and external radiation dose, whereas the results from the case-control study did not find any significant association.

The matching on date of birth and date of entry meant that workers of the same age were working at the same times. This seems to have had the effect that workers in the same matched set have broadly similar recorded doses. The apparent over-matching on date of entry has distorted the parameter estimate of the risk of leukemia on cumulative dose by introducing matching on dose.

In the analysis phase, stratification is a strategy that analyzes data and presents results according to subgroups of patients, or “strata,” of similar characteristics, if there is a potential confounder. It ensures that only cases and controls with similar levels of a potential confounding variable are compared. Subjects are divided into strata or subgroups depending on the level of a potential confounder, for example, age. Considering age as a confounder in a study on diffusion tensor MRI of cartilage and stratifying the sample by confounder the investigators should look at anisotropy of cartilage within two or more different pediatric age groups (strata) (5 to 10, 11 to 14, and 15 to 18 years) separately.

The relationship between the variable of interest and the outcome is then separately analyzed for each stratum. The major problem with stratification is sample size and power, since this strategy decreases the sample by subdividing it further, resulting in lower likelihood of demonstrating a significant association when present if the study sample size is inappropriate (type II error).

Regression Adjustment

Regression adjustment is a strategy that exists because the relationships among variables can be complex, with variables being related to one
another as well as to the outcome of interest. In addition, the effect of one variable may be modified by the presence of others, and the joint effects of two or more variables can be greater than the sum of their individual effects. The rationale for the utilization of this strategy (“adjusting for the confounder[s]”) is to investigate potential confounder variable(s) and its (their) association with the outcome, integrating this information into the analysis, which is done using multivariate regression models.14

**Effect Modifier**

Whereas a confounder is associated with both exposure and outcome, an effect modifier is a variable that—although not associated with the exposure—is associated with the outcome and differentially (positively or negatively) modifies the observed effect of an exposure (risk factor or treatment) on the outcome (disease status) (Fig. 5.5).15 Different groups have different risk estimates when effect modification is present and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion. For example, risk of malignancy due to exposure to ionizing radiation may be different for men and women, as well age at first exposure may impact the risk too. A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion. For example, risk of malignancy due to exposure to ionizing radiation may be different for men and women, as well age at first exposure may impact the risk too. A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion.

A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion. For example, risk of malignancy due to exposure to ionizing radiation may be different for men and women, as well age at first exposure may impact the risk too. A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion.

A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion. For example, risk of malignancy due to exposure to ionizing radiation may be different for men and women, as well age at first exposure may impact the risk too. A cohort study examining risks of radiation exposure may yield incorrect estimates of the effect of radiation exposure if the impact of effect modification is not accounted for. Since effect modification is generally considered to be a biological phenomenon, it cannot be prevented in the study design phase and, therefore, an estimate of the impact of the exposure on outcome for the entire sample may lead to an incorrect conclusion.

**Reporting Observational Studies**

The analysis of observational studies is covered in Chapter 13. It is important to understand that the specific statistics that one gets following an observational study will depend on the type of design. For example, relative risk (RR) is a common statistic that is calculated as part of a cohort study. On the other hand, following a case-control study, one generally computes and reports an odds ratio (OR). Although both RR and OR quantify how strongly the presence or absence of exposure is associated with the presence or absence of outcome in a given population, the same statistic cannot be calculated for each study design. The reason for this is that in a case-control study, the groups (cases and controls) are determined by the presence of the outcome and by the ratio of these by investigator. Therefore, incidence rates of outcomes for a given exposure cannot be calculated. On the other hand, in a cohort study, since groups are determined by exposure, incidence of outcome by group can be computed.

While OR and RRs are not the only outcome measures for nonexperimental studies, they are the most frequently used. It is important to consult with a statistician when analyzing and reporting results of observational studies, to ensure that you are using the appropriate measures for your study design that most clearly and accurately report the results of your study. It is suggested that you consult with a statistician early in the design of your study to ensure that the study is well designed and that the optimal data is collected in a format that will allow you to answer the question of your study.

**STROBE Statement**

Unfortunately, historically observational studies were poorly reported. In order to address this issue, a group of methodologists, researchers, and editors developed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)19 to improve the quality of reporting of studies. The STROBE statement contains the aspects of observational trials that should be reported (Fig. 5.6). Standardization of the reporting of observational studies makes it easier to appraise the results of these studies, and to combine results of studies in systematic...

Study Design

This study was a retrospective cohort study designed to compare utilization and spending between different types of noninvasive cardiac tests in a Medicare population. The authors compared functional (stress echocardiography, stress myocardial perfusion scintigraphy [MPS], and exercise electrocardiogram) testing with anatomical (coronary CT angiography [CCTA]) testing.
## TITLE and ABSTRACT

<table>
<thead>
<tr>
<th>Item number</th>
<th>Recommendation</th>
</tr>
</thead>
</table>
| 1           | (a) Indicate the study’s design with a commonly used term in the title or the abstract  
(b) Provide in the abstract an informative and balanced summary of what was done and what was found |

## INTRODUCTION

- **Background/rationale**
  - 2 Explain the scientific background and rationale for the investigation being reported
- **Objectives**
  - 3 State specific objectives, including any prespecified hypotheses

## METHODS

### Study design

- **Participants**
  - 4 Present key elements of study design early in the paper
  
- 5 Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
  
- (a) Cohort study—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up  
- (b) Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls  
- (c) Cross-sectional study—Give the eligibility criteria, and the sources and methods of selection of participants
  
- (b) Indicate the number of participants with missing data for each variable of interest
  
- (c) Explain how the study size was arrived at

### Variables

- **Data sources/measurement**
  - 7 Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable

- **Bias**
  - 8* Describe comparability of assessment methods if there is more than one group

- **Study size**
  - 9 Describe any efforts to address potential sources of bias

- **Statistical methods**
  - 10 Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why

- **RESULTS**

- **Participants**
  - 13* (a) Report the numbers of individuals at each stage of the study—e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed  
  
- (b) Give reasons for non-participation at each stage

- **Descriptive data**
  - 14* (a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders

- (b) Indicate the number of participants with missing data for each variable of interest

- (c) Cohort study—Summarise follow-up time (e.g., average and total amount)

- **Outcome data**
  - 15* Cohort study—Report numbers of outcome events or summary measures over time  
  
  Case-control study—Report numbers in each exposure category, or summary measures of exposure  
  
  Cross-sectional study—Report numbers of outcome events or summary measures

- **Main results**
  - 16 (a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included

- (b) Report category boundaries when continuous variables were categorized

- (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period

- **Other analyses**
  - 17 Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses

## DISCUSSION

- **Key results**
  - 18 Summarise key results with reference to study objectives

- **Limitations**
  - 19 Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias

- **Interpretation**
  - 20 Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence

- **Generalisability**
  - 21 Discuss the generalisability (external validity) of the study results

## OTHER INFORMATION

- **Funding**
  - 22 Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

---

* Give such information separately for cases and controls in case-control studies, and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

**Fig. 5.6** The STROBE statement. Checklist of items that should be addressed in reports of observational studies: (a) introduction and methods; (b) results, discussion, and other information. (Reproduced under agreement with the *Journal of Clinical Epidemiology.*)
We will review the key aspects of the study using the STROBE statement as a guideline tool for critical appraisal.

**Title and Abstract**

1a. Title: Yes: Page 2128: The study design is reported in the abstract.

**Introduction**

2. Background/rationale: Yes: Pages 2128, 2129: Rationale is explained.
3. Objectives: Page 2129: Yes: Objective is reported.

**Methods**

4. Study design: Yes: Page 2129: Key elements of study design are reported in the abstract, data sources, patient cohort, and episode of care sections. Exposures: CCTA or stress testing. Outcomes: Cardiac catheterization, coronary revascularization, acute myocardial infarction, all-cause mortality, and total and CAD-related Medicare spending over 180 days of follow-up.
5. Setting: Yes: Page 2129: Setting details are reported.
6. Participants: Yes: Page 2129: The authors explain that the cohort was drawn from a 20% random sample of Medicare beneficiaries and how they were able to track down users and Medicare codes.
7. Variables: Yes: Pages 2128, 2129, 2130: Exposures and outcomes are reported in the abstract and methods. Potential confounders are reported in the statistical analysis subsection.
8. Data sources/measurements: Yes: Pages 2129, 2130: The authors reported data sources (page 2129) and how differences between groups were assessed (statistical analysis, page 2130).
9. Bias: Yes: Page 2130: The regression models used in the study controlled for potential confounders. The models also included year and month dummies to capture trends over time.
10. Study size: No: Page 2129: The authors do not explain the rationale for using a 20% random sample of traditional Medicare beneficiaries.
11. Quantitative variables: Yes: Page 2130: In the statistical analysis section, the authors explained in detail how quantitative variables were handled in the analyses.
12. Statistical methods: Partial reporting: Page 2130. The authors:
   a) Yes: Described all statistical methods, including those used to control for confounding.
   b) Yes: Used joint tests across all four groups.
   c) No: Did not mention whether there were missing data in the study and, if there were, how they were addressed. In the limitations section they referred to lack of data on clinical symptoms or test findings, but this seems to be overall, not related to a certain number of cases of the study sample.
   d) No: Used administrative data; there was no information about missing data.
   e) Yes: Reported sensitivity analyses using a shorter follow-up time (90 days).

**Results**

13. Participants:
   a) Yes: Reported in Table 1, page 2131.
   b) Not applicable: The authors used administrative data.
   c) Yes: Page 2130: Study flow is reported.
14. Descriptive data:
   a) Yes: Reported in Table 1, page 2131.
   b) No.
   c) Partial: The authors reported in Figure 2 (page 2133) unadjusted 180-day rates (rather than values at individual time points) of catheterization and revascularization after CCTA or MPS, and selected Medicare beneficiary characteristics.
15. Outcome data: Yes: The authors reported in Tables 2 and 3, pages 2132 and 2134, adjusted associations between noninvasive cardiac testing and 180-day clinical outcomes/spending in Medicare beneficiaries, 2006–2008.
16. Main results:
   a) Not applicable: The authors used administrative data.
   b) Not applicable: The authors used administrative data.
   c) Yes: Study flow is reported. Most results are available in Tables 2 and 3 and Figure 2.
17. Other analyses: Yes: Pages reported sensitivity analyses and results of joint tests across all four groups.

Discussion
19. Limitations: Yes: Limitations are reported on page 2135.
21. Generalizability: Yes: Discussed on page 2135: “because it is based on a sample of the Medicare population, it also is likely more representative and generalizable to the nearly 40 million individuals older than 65 years in the United States.”
22. Other information: Yes: Page 2134: Received funding was reported.

<table>
<thead>
<tr>
<th>Highlights of Key Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Nonexperimental methods are approaches whereby the investigator does not alter or manipulate the circumstances of the participants in the research study.</td>
</tr>
<tr>
<td>• The major types of nonexperimental designs include case series, cohort studies, and case-control studies.</td>
</tr>
<tr>
<td>• The completeness of the data afforded by a prospective study design need to be balanced against the additional time and cost of this approach when compared to a retrospective design.</td>
</tr>
<tr>
<td>• Nonexperimental designs are subject to confounding, which is the relationship of another variable to both exposure and outcome, not on the causal pathway, which affects the outcome.</td>
</tr>
<tr>
<td>• Estimates of association between exposure and outcome in case-control studies are expressed by the odds ratio, whereas the relative risk is used in cohort studies.</td>
</tr>
</tbody>
</table>

References
Randomized Controlled Trials

Ivan R. Diamond and Brian M. Feldman

Learning Objectives

- To understand the distinction between experimental and nonexperimental designs.
- To understand what is meant by bias and confounding.
- To understand the key components of a randomized controlled trial.
- To understand the rationale for randomization.
- To understand the rationale for blinding in randomized controlled trials.
- To learn how to critically appraise the results of a randomized controlled trial.

Concepts

Experimental Design

The randomized controlled trial is a true experiment as the investigator manipulates the intervention received by the subject in a controlled manner. The intervention received is also determined by a random process and not by other factors. These attributes allow for optimal assessment of the causal relationship between the intervention and the outcome. This concept is known as internal validity.

Consider an alternative situation whereby the investigator defines a number of treatment paths for a particular disorder. However, the choice of treatment is up to the clinician caring for the patient. At the conclusion of the study, the investigator compares the results of the various treatments. This is a quasi-experimental design with lower internal validity than the randomized trial. The reason is that if there turns out to be a difference between treatments, it may be related to the factors that influenced the choice of treatment received rather than the treatment itself. This concept, known as confounding by indication, has already been discussed in Chapter 5 as well as bias, and is discussed in more detail later in this chapter. While quasi-experimental designs may be used to assess the causal relationship between treatment and outcome, the level of inference (i.e., internal validity) is lower than for a randomized design.

Bias and Confounding

Bias

While the goal of research is to learn the “truth,” bias refers to any factor that causes the results of a study to deviate from the “truth.” Bias does not imply malice, rather it is generally the result of a systematic error in research methodology. It is often difficult to predict the impact that bias may have on the results of a study. Bias may not be able to be addressed statistically. It is critical to understand the potential sources of bias when designing a randomized controlled trial in order to avoid them. We consider some of the more common forms of bias that affect interventional studies and later highlight aspects of the randomized design that attempt to address these issues.

Selection Bias

Selection bias, also known as sampling bias, occurs when enrolling subjects into the trial. Selection bias occurs when only a limited subset of potentially eligible subjects is enrolled in the trial. While it is never possible to enroll all potentially eligible subjects in a trial, selection bias occurs when there is an important difference between those who are enrolled and the overall group of those who are...
Randomized Controlled Trials and meta-analyses may result in biased (typically overly optimistic) estimates of a treatment effect or diagnostic accuracy of imaging tests. In order to limit this bias, most journals require, prior to consideration for publication, that the trial be registered at its outset on a public trial register such as www.clinicaltrials.gov. Therefore, even if negative trials are never published, authors of reviews will be aware of their existence and can contact the investigators for results.

Moreover, in the conduct of a review, restriction of language of manuscripts to English simplifies its performance but reduces the generalizability of the review results to non-English-speaking countries.

Confounding

A confounding variable is a factor that distorts the true relationship of the study variable of interest because it is also related to the outcome of interest. Confounding occurs when the impact of an intervention is altered because of the association of the intervention with other factors that influence the outcome. The confounding variable may mask an actual association or it may falsely demonstrate an apparent association between the study variables where no real association between them exists. One of the most prevalent types of confounding is confounding by indication.

Confounding by indication refers to the situation whereby patients with a particular prognosis are allocated preferentially to a particular treatment or intervention arm. These patients are systematically different from those who receive alternate treatments. This systematic difference leads to apparent differences in outcome between the treatments or interventions. However, the observed differences are due to the confounders rather than the therapies or interventions received.

Fig. 6.1 illustrates confounding by indication in a hypothetical nonrandomized study that examined survival of patients undergoing endovascular compared to open aortic aneurism repair. The results demonstrated that 5-year post-procedure survival in the endovascular group was worse than with open repair. However, when you look closer...
Research Methods in Radiology

at the data you discover that older and sicker patients underwent the endovascular procedure, whereas younger patients with less comorbidity underwent open repair. Since both age and comorbidity influence overall survival (the outcome), independently of the procedure performed, they are confounders. In this example, it is possible that the difference in the outcomes between the treatments is related to these confounders rather than inherent differences in the procedures.

It is difficult to predict the impact of various confounders, which may exaggerate or diminish treatment effects. As well, it is impossible to identify all confounders. Therefore, even if an investigator is able to identify and account for confounders in the research design, there will always be unknown confounders that may influence the results.

The main strength of the randomized trial is that the random assignment to treatment groups leads to a balanced distribution of both known and unknown confounders across treatment groups. As a result, the randomized controlled trial is able to assess the impact of a particular intervention while minimizing the influence of confounders.

Overview of a Randomized Controlled Trial

Fig. 6.2 presents a basic outline of a randomized trial. The trial depicted in this diagram is the simplest type of randomized design, the parallel group design. In the parallel group randomized controlled trial, there is a single experimental intervention that is compared to a single control intervention. Busse et al. provided an overview of more complex trial designs. The key components of the trial—study sample, randomization, intervention,
Randomized Controlled Trials

Recruit. However, the more dissimilar subjects are from one another, the more heterogeneous the treatment response will be. Increasing heterogeneity results in a need for a larger number of subjects in order to answer the research question. Finally, investigators need to consider subject safety in developing their eligibility criteria.

The choice of eligibility criteria has important implications when applying the results of the trial. The more similar the trial subjects are to a prospective population of patients, the more likely that the results will be applicable to those patients. This concept is known as generalizability or external validity. When applying the results of a randomized trial, it is important to not only look at the eligibility criteria as described in the methods section of the paper, but also the characteristics of the subjects who actually participated in the trial. If there are important differences in the characteristics of participants compared with the target population, then the trial may have been subject to a selection bias that may alter the conclusions of the trial.

Randomization

Randomization is the process of assigning subjects to groups based on chance. Generation of the Randomization Sequence

Generation of the randomization sequence refers to developing the order to which subjects will be assigned to treatment groups. There are a number of methods for randomization, but regardless of the method used, it is essential that subjects are assigned to treatment groups based on a random process and not another, potentially predictable, factor.

For example, some trials have used hospital file numbers, birth date, or day of the week to assign subjects to intervention groups. While the relationship between these factors and the patient is random, since these factors are known prior to enrollment they may influence the decision to enroll a patient into the trial. If this were to occur, the trial is subject to allocation bias, which is a subset of selection bias.

Study Sample

The study sample refers to the subjects who participate in the trial. A key feature of a randomized controlled trial is that all subjects must be able to receive any of the interventions that are being studied as part of the trial. In order to do this, researchers develop eligibility (inclusion and exclusion) criteria.

When developing eligibility criteria, there are a number of factors that the investigators need to balance. First, the investigators need to decide who the target population for their trial will be. This is typically the patient population to whom they would like the results of the trial to apply. Second, investigators need to consider how feasible it will be to recruit subjects who meet their eligibility criteria. If the trial is being performed in a rare disease, then the investigator will need to consider enrolling subjects at multiple sites in order to meet their recruitment target within a reasonable period of time. Usually the broader the eligibility criteria, the easier it will be to recruit. However, the more dissimilar subjects are from one other, the more heterogeneous the treatment response will be. Increasing heterogeneity results in a need for a larger number of subjects in order to answer the research question. Finally, investigators need to consider subject safety in developing their eligibility criteria.

The choice of eligibility criteria has important implications when applying the results of the trial. The more similar the trial subjects are to a prospective population of patients, the more likely that the results will be applicable to those patients. This concept is known as generalizability or external validity. When applying the results of a randomized trial, it is important to not only look at the eligibility criteria as described in the methods section of the paper, but also the characteristics of the subjects who actually participated in the trial. If there are important differences in the characteristics of participants compared with the target population, then the trial may have been subject to a selection bias that may alter the conclusions of the trial.
Appropriate methods for randomization range from low-tech approaches such as dice or random number tables to sophisticated computer-based algorithms.\textsuperscript{10} Generation of the sequence to which patients will be assigned to the trial is often done prior to the start of the trial but may be done on an ongoing basis when computer-based approaches are used.

**Block Randomization**

When completely random methods, such as a random number table, are used to generate the randomization sequence, the treatment group sizes may be quite unequal in smaller trials just due to chance. Block randomization is a technique of randomization used in smaller studies to ensure more equal distribution of subjects across treatments. Blocking is achieved by creating blocks of a size that are multiples of the total number of treatments (e.g., AABB, BAAB, or BACACB, BACBCACAB). In each block, there is an equal number of subjects in each group but the order of treatments in the block is determined randomly. Therefore, after each block is complete, there will be equal numbers of patients in the treatment groups.

While blocking seems to be a good solution of the problem of unequal group sizes, the problem with blocking is that since the randomization sequence is broken into small components that all have an equal number of subjects from each group; it may become possible to predict the next subjects toward the end of a single block. Strategies to avoid this include varying block sizes and keeping the block size concealed from the investigator.

**Stratification**

The goal of randomization is to ensure that confounders are randomly distributed across the treatment groups. This does not necessarily mean that the groups will have an equal number of subjects with specific characteristics although as trials become larger, then the likelihood of this increases. Sometimes, investigators will want to ensure that they have sufficient numbers of subjects with particular characteristics across the treatment groups. This may be done if the investigators are planning on performing subgroup analyses using these characteristics or if they want to ensure that each treatment group represents the spectrum of disease. Stratification is a process whereby subjects are randomized and allocated to groups based on key characteristics to ensure that there are sufficiently large subgroups of patients with particular characteristics so that particular issues can be addressed with sufficient power.

The simplest approach to stratification would be to develop different group assignments for each combination of subject characteristics—known as strata. For example, if the investigators wished to stratify by gender and disease severity, they would require four randomization tables (male/mild, male/severe, female/mild, female/severe). The major issue with stratification is that the number of sequences needed increases exponentially with the number of stratification variables.\textsuperscript{11} Stratification leads to methodological complexity as each combination of strata forms essentially a separate randomized trial that needs to be accounted for in the analysis. But as the number of strata increases, the number of subjects within each stratum decreases for a fixed trial size and it is possible that some strata will have only treated or untreated patients. This also has implications for subgroup analyses as the study will become underpowered for these analyses because of the low number of subjects within the various strata.\textsuperscript{11} For this reason, one should generally not stratify on more than two variables, and in larger trials, stratification may not be necessary.

**Allocation Concealment**

Regardless of the method used to generate the randomization sequence, it is essential that a potential subject or investigator not be able to know or predict the group assignment prior to enrollment in the trial.\textsuperscript{12} Such knowledge may influence who participates in the study which may introduce allocation bias whereby different patients may be enrolled or not enrolled depending on the knowledge to which group they would be assigned.
Allocation concealment refers to the process of keeping knowledge of the randomization sequence known from investigators and potential subjects prior to enrollment. There are a number of methods that can be used to achieve allocation concealment. Allocation concealment is inherent in computer-based systems where randomization occurs at the time of enrollment. Alternatively, investigators may choose to employ central randomization whereby a third party, such as a pharmacy or central office, holds the randomization sequence, releasing the group assignment only once the subject is enrolled in the trial.

While central randomization may seem to be ideal, it can be costly especially for trials that enroll subjects at unusual hours of the day. Envelopes are used by many investigators as they are a simple way to reveal the group assignment following enrollment. However, use of envelopes is controversial as an approach to allocation concealment. If envelopes are used, it is essential that they are opaque, sealed, and sequentially numbered and cannot be resealed once opened. This ensures that there is no way to determine group assignment until the envelope is opened. Also, once the envelope is opened the subject is considered to be enrolled and assigned to a particular group and another envelope cannot be chosen.

The importance of allocation concealment was highlighted by Shultz et al., who reviewed the results from a number of meta-analyses looking at impact of methodological issues on estimates of treatment response. In trials where allocation concealment was not performed, the trial overestimated the treatment effect by 40%. In trials where it was unclear how allocation concealment was achieved, treatment estimates were 30% higher. Therefore, it is essential that randomized trials have adequate allocation concealment procedures and that these procedures are described in the publications reporting the methods and trial results.

**Allocation Concealment versus Blinding**

While allocation concealment protects from knowledge of group assignment prior to enrollment in the trial, blinding keeps the group assignment secret after enrollment has occurred. Blinding usually lasts for the duration of the trial, although some trials maintain the blinded group status until after the trial has been analyzed. Blinding limits ascertainment bias, which occurs when knowledge of group assignment influences the delivery of the intervention or assessment of outcomes. Controlling for other factors, when compared with blinded trials, unblinded trials overestimated treatment response by 17%.

While it may seem optimal to blind all randomized trials, it is not always feasible. For example, it may not be possible to blind patients undergoing a trial of an image-guided versus a surgical procedure. However, it may still be possible to blind outcome assessors. Preserving blinding may also be ethically challenging, such as the notion of using a sham procedure in a trial comparing an invasive procedure to a noninvasive procedure.

The decision to blind should be based on the type and magnitude of bias that would be introduced if the trial were not blinded as well as the feasibility of blinding. This depends on the specific intervention that is being studied as part of the trial and also the outcome measures used. In general, subjective outcomes are more prone than objective outcomes to bias from lack of blinding.

Blinding can be done for different groups involved in the trial, namely, participants, caregivers, outcome assessors, and statisticians. Frequently one sees the terms single, double, or multiple blind. Single blind refers to one of the groups being blinded; double, two groups—usually subjects and investigators. Multiple blind refers to multiple groups. Since these terms are not useful for a reader in knowing exactly who was blinded as part of the study, it is preferable to describe exactly which groups were blinded and how rather than using these terms.

**Intervention**

The intervention refers to the treatment, procedure, or drug being evaluated in the trial. The intervention is compared with the “control” treatment which may be another.
medication, modality, procedure, nonsurgical approach, or placebo.\textsuperscript{15} When designing a trial the intervention needs to be clearly defined and described so that it is clear exactly what was done, partly so that the trial can be replicated by others.

**Equipoise**

The choice of interventions to be studied in a randomized controlled trial is typically based on equipoise.\textsuperscript{16} Equipoise classically is defined as expert opinion that the treatments being considered in the trial have equivalent efficacy. Given that the classic definition of equipoise is quite stringent, community equipoise or uncertainty has been suggested as the basis for a trial. By this alternate definition, trials are justified when there is a difference of expert opinion with regard to the relative merits of the treatments or experts are uncertain which treatments are best. Since health care cannot be compromised by participation in a trial, all treatments must have sound rationale for their use and treatments with good evidence of benefit or harm relative to another cannot be compared to one another in a randomized trial.

**Outcomes**

Outcomes refer to the measures taken at the end of the trial that will be used to assess whether the intervention and control groups differ (i.e., that one treatment is better than the other). The choice of outcome depends on the objectives of the study.\textsuperscript{17} There are a number of types of outcome measures including clinical outcomes (e.g., recurrence of stroke, length of stay in hospital), biologic outcomes (e.g., hemoglobin), patient-derived outcomes (e.g., quality of life or satisfaction), and economic outcomes (e.g., cost). While trials will typically have a number of outcomes, one is designated as the primary outcome. This should be the outcome that addresses the primary question that the trial aims to answer. This is also the outcome on which sample size calculations are based.

Whatever the outcome measure used, it is essential to adequately define it. The importance of this can be demonstrated from a study of a number of definitions of surgical site infection.\textsuperscript{18} Depending on the definition used, the overall infection rate varied from 7% to 20%. When choosing an outcome measure and its definition, it is important to consider the implications that this will have when applying the results of the trial. For example, the observed complication rates will be most accurate in predicting risk of complications outside the trial, if they are defined the same way as the trial defined them.

### Analyzing and Reporting Clinical Trials

Clinical trials can be challenging studies to perform given their cost and complexity. It is mandatory that investigators collaborate with a statistical and methodological expert throughout the trial to ensure that the trial is able to achieve its stated objectives. This is particularly important during the design and analysis phase of the trial.

**Intention-to-Treat Analysis**

Subjects in clinical trials may drop out from the trial, or receive the treatment that is being used in the other group (known as “cross-over”). Dropouts and cross-overs pose a challenge when analyzing the results of a randomized trial as reasons for crossing over or dropping out are generally not random.\textsuperscript{19} As a result of this, treatment groups that were comparable at the start of the trial (because of randomization) will be systematically different from one another at the end of the trial due to dropouts and cross-overs. Removing noncompliers may destroy the unbiased comparison provided by randomization. Unless appropriate analytic approaches are used, bias will be introduced.

The intention-to-treat analysis is a conservative strategy for analysis of randomized trials and, therefore, the preferred method for dealing with dropouts and patients who cross over.\textsuperscript{19} Subjects are analyzed according to the group to which they were originally randomized regardless of the treatment they
have received. Using this strategy, investigators include all patients at the study entry, regardless of whether they received the treatment or intervention to which they were randomly allocated, subsequently withdrew from the study, or deviated from the study protocol. Two alternatives to intention-to-treat are the “as-treated” analysis, where patients are classified according to the treatment actually received, and the per protocol analysis, where patients are included only if their treatment went according to study protocol (for dose, timing, compliance, etc.).

The intention-to-treat principle maintains the benefits of randomization minimizing the influence of withdrawals, noncompliers, and patients lost to follow-up and provides an estimate of treatment effectiveness that more closely reflects the “real-world” effect of prescribing a treatment. An advantage of this strategy is that it allows greater generalizability of study results and minimizes the risk of a type I error (“to say that there is a difference between groups when in reality there is not”) because of its greater caution. A limitation of this strategy is that such an analysis is less likely to show a positive treatment or intervention effect, especially in studies that randomize patients who have little chance of benefiting from the intervention. For example, in a drug trial, where patients may switch drugs due to side effects, the intention-to-treat analysis will provide an estimate of the overall effectiveness of the drug that takes into account the fact that not all patients will be able to take the drug.

Fig. 6.3 illustrates a hypothetical example of how excluding patients who do not receive the intervention to which they were originally assigned can introduce bias. Imagine a randomized trial of 200 patients who require the use of peripherally inserted central catheters (PICCs) for different clinical indications; 100 are assigned to receive cuffed PICCs and the other 100 to receive uncuffed PICCs. As per the protocol, the PICC is to stay in for a minimum period of 1 week and the primary outcome is the complication rate for PICCs at the end of this period. Let’s assume that the investigators expect uncuffed PICCs to have an equal or a higher rate of complication (infection, malposition, thrombus formation).
as compared to cuffed PICCs. In the cuffed PICC arm, 20 out of 100 patients developed a complication, following the 1-week period. In the uncuffed PICC arm, 10 patients withdrew prior to the first week, and of the 90 patients remaining on the trial, 10 developed a complication following the 1-week period. In this example if we restrict the analysis in the uncuffed PICC arm to the patients who had the PICC for the minimum period of 1 week (a per protocol analysis), the event (complication) rate will be 11% (10/90); however, the rate in the cuffed arm will be 20% (20/100). Since the investigators had assumed that uncuffed PICCs would present with an equal or higher rate of complications, these values represent a spurious (unexpected) result. Alternatively, if we include all randomized patients, according to the intention-to-treat principle, and take a worst-case approach of assuming a complication in those who ended the trial early, we see that there were 20 events in each arm and no impact of cuffing on complications.

While the intention-to-treat analysis should be done as the primary analysis as it provides the least biased estimate of treatment response, an as-treated or per protocol analysis is often included as a secondary analysis.

Reporting Clinical Trials – CONSORT
Randomized trials are often poorly reported. Therefore, a group of biostatisticians and clinical epidemiologists developed the CONSORT statement, which is a set of guidelines for the reporting of results from randomized trials. The CONSORT guidelines include all aspects of the trial that should be reported (Fig. 6.4). This allows for critical appraisal of the trial results. Standardized reporting also allows for combining results across multiple trials in meta-analyses. Most journals require that reports from clinical trials conform to CONSORT in order to be considered for publication.

Example

This study was a randomized controlled trial designed to compare the change in heart rate between low osmolar contrast agents versus standard osmolar contrast agents in those undergoing CT pulmonary angiography. The rationale for the study is that heart rate affects image quality. We will review the key aspects of the study design in order to illustrate the concepts discussed in this chapter.

Study Design
The study was a parallel group randomized controlled trial in patients undergoing CT angiography. The trial was performed at multiple centers. Patients and investigators were blinded to the type of contrast medium used.

Study Sample
Subjects were 130 patients > 18 years of age undergoing contrast-enhanced CT of the chest for suspected pulmonary embolism. The authors adopted a simple approach to their choice of subjects, using local standard of care as their criterion for suspected pulmonary embolism. This may lead to differences in subject characteristics between centers, as local standard of care may differ. This may lead to a difference in the likelihood of pulmonary embolism at various sites. However, unless the cardiovascular response to contrast depends on the presence of absence of pulmonary embolism, these differences should not negatively affect the trial. More restrictive inclusion criteria may have made the trial more challenging to perform. The exclusion criteria primarily dealt with safety by excluding patients with severe congestive heart failure or other medical conditions. Women of childbearing age and those who had a pacemaker or defibrillator were also excluded.

Randomization
Since this was a small trial, block randomization using blocks of four was used. Block size did not vary. The randomization sequence was computer-generated. Further details
<table>
<thead>
<tr>
<th>Section/Topic</th>
<th>Item No</th>
<th>Checklist Item</th>
<th>Reported on page No</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title and abstract</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1a</td>
<td>Identification as a randomised trial in the title</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b</td>
<td>Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)</td>
<td></td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a</td>
<td>Scientific background and explanation of rationale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2b</td>
<td>Specific objectives or hypotheses</td>
<td></td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>Description of trial design (such as parallel, factorial) including allocation ratio</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3b</td>
<td>Important changes to methods after trial commencement (such as eligibility criteria), with reasons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>Eligibility criteria for participants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>Settings and locations where the data were collected</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>The interventions for each group with sufficient details to allow replication, including how and when they were actually administered</td>
<td></td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6a</td>
<td>Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6b</td>
<td>Any changes to trial outcomes after the trial commenced, with reasons</td>
<td></td>
</tr>
<tr>
<td><strong>Sample size</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7a</td>
<td>How sample size was determined</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7b</td>
<td>When applicable, explanation of any interim analyses and stopping guidelines</td>
<td></td>
</tr>
<tr>
<td><strong>Randomisation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8a</td>
<td>Method used to generate the random allocation sequence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8b</td>
<td>Type of randomisation; details of any restriction (such as blocking and block size)</td>
<td></td>
</tr>
<tr>
<td><strong>Allocation concealment mechanism</strong></td>
<td>9</td>
<td>Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned</td>
<td></td>
</tr>
<tr>
<td><strong>Implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions</td>
<td></td>
</tr>
<tr>
<td><strong>Blinding</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11a</td>
<td>If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11b</td>
<td>If relevant, description of the similarity of interventions</td>
<td></td>
</tr>
<tr>
<td><strong>Statistical methods</strong></td>
<td>12a</td>
<td>Statistical methods used to compare groups for primary and secondary outcomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12b</td>
<td>Methods for additional analyses, such as subgroup analyses and adjusted analyses</td>
<td></td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13a</td>
<td>For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13b</td>
<td>For each group, losses and exclusions after randomisation, together with reasons</td>
<td></td>
</tr>
<tr>
<td><strong>Recruitment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14a</td>
<td>Dates defining the periods of recruitment and follow-up</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14b</td>
<td>Why the trial ended or was stopped</td>
<td></td>
</tr>
<tr>
<td><strong>Baseline data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>A table showing baseline demographic and clinical characteristics for each group</td>
<td></td>
</tr>
<tr>
<td><strong>Numbers analysed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups</td>
<td></td>
</tr>
<tr>
<td><strong>Outcomes and estimation</strong></td>
<td>17a</td>
<td>For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17b</td>
<td>For binary outcomes, presentation of both absolute and relative effect sizes is recommended</td>
<td></td>
</tr>
<tr>
<td><strong>Ancillary analyses</strong></td>
<td>18</td>
<td>Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory</td>
<td></td>
</tr>
<tr>
<td><strong>Harms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)</td>
<td></td>
</tr>
<tr>
<td><strong>Discussion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses</td>
<td></td>
</tr>
<tr>
<td><strong>Generalisability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>Generalisability (external validity, applicability) of the trial findings</td>
<td></td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence</td>
<td></td>
</tr>
<tr>
<td><strong>Other information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>Registration number and name of trial registry</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Where the full trial protocol can be accessed, if available</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>Sources of funding and other support (such as supply of drugs), role of funders</td>
<td></td>
</tr>
</tbody>
</table>

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming for those and for up to date references relevant to this checklist, see [www.consort-statement.org](http://www.consort-statement.org).*

Fig. 6.4 CONSORT checklist. (Reproduced under the Creative Commons Attribution License. Checklist downloaded from [http://www.consort-statement.org](http://www.consort-statement.org). Accessed on January 23, 2014.)
regarding allocation concealment were not provided in the manuscript, which is a weakness in the reporting of the trial, and perhaps in the trial methodology.

**Blinding**

Blinding was achieved by ensuring that preparation of the contrast media was done by a technologist, nurse, or physician who was not involved in study evaluations. Since low and iso-osmolar contrast media are identical in appearance, it was straightforward to blind in similar syringes.

**Intervention**

The intervention involved the administration of identical doses of a low osmolar versus a regular osmolar contrast agent. The authors describe a specific protocol for the injection of the contrast media as well as a detailed description of the imaging protocol. This allows for readers to critically appraise the methodology as well as to replicate the study if needed. If the results of the trial are combined in a meta-analysis, it would be important to refer to the methodology, so that one can ensure that methodologically similar trials are combined.

**Outcomes**

The outcome of the trial was the change in heart rate from the beginning to the end of the CT examination. Precise details for measuring the heart rate are provided in the paper to allow for reproducibility. In addition to recording heart rate prior to, during, and after the procedure, the investigators also captured a rhythm strip so that the maximum heart rate could be recorded. It is interesting that a measure of image quality was not included as an outcome, as that is the clinically important outcome. The reason for this is that image quality may be difficult to measure and quantify. In this case, heart rate can be termed a surrogate outcome.

**Analysis**

Although there were patients who were randomized who did not receive their assigned treatment, the paper does not state whether intention-to-treat analysis was performed. This is a further weakness in reporting or perhaps in trial methodology.

**Highlights of Key Points**

- Randomized controlled trials are a true experimental design.
- Bias is a systematic error in research methodology that causes the results of a study to deviate from the “truth.”
- Randomization ensures known and unknown confounders are randomly distributed across treatment groups. Randomization limits confounding by indication, allowing for more valid assessment of the impact of a particular intervention.
- Allocation concealment is a key aspect of a randomized controlled trial and protects from allocation bias, a form of selection bias.
- While blinding may not be feasible for all aspects of all trials, the rationale for blinding randomized controlled trials is to limit ascertainment bias during administration of interventions or during outcome assessment.
- Registration of randomized trials at trial onset is required by most biomedical journals. The goal of registration is to limit publication bias.
- It is mandatory that investigators collaborate with a statistical and methodological expert throughout the trial to ensure that the trial is able to achieve the trial objectives.
- Analysis by intention-to-treat is the preferred analytic method to limit the bias that results from subjects who drop out from or cross-over during a randomized trial.
- CONSORT is a set of guidelines for the high-quality reporting of the methods and results of randomized controlled trials. Compliance with CONSORT is mandated by most biomedical journals.
References

2. Dickersin K. The existence of publication bias and risk factors for its outcome. JAMA 1990;263:1385–1389
11. Therneau T. How many stratification factors are “too many” to use in a randomization plan? Control Clin Trials 1993;14:98–108
7 Systematic Reviews, Evidence-Based Imaging, and Knowledge Translation

Andrea S. Doria, Jennifer Stinson, and Prakeshkumar Shah

1973: “It is surely a great criticism of our profession that we have not organized a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomized controlled trials.”

—Professor Archibald Cochrane (1909–1988)

■ Learning Objectives

- To provide basic concepts on definitions of unstructured reviews, systematic reviews, meta-analyses, pooled analyses, evidence-based imaging, and guidelines.
- To outline and describe the steps for conducting a systematic review and meta-analysis and for developing practical clinical guidelines based on evidence derived from available systematic reviews/meta-analyses.
- To discuss the available tools for assessing the quality of reporting and methodology of papers and systematic reviews in diagnostic imaging.
- To introduce concepts on implementation and knowledge translation, T1 and T2 translational research, key guideline questions for knowledge translation activities, and the role of this new science on accelerating the dissemination of knowledge in the radiological sciences.

■ Introduction

Evidence-based medicine is “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external evidence from systematic research.”

Evidence-based imaging (EBI), in contrast to the traditional paradigm, acknowledges that intuition, unsystematic clinical experience, and pathophysiologic rationale are insufficient grounds for clinical decision making, and stresses the examination of evidence from clinical research in a critical manner. EBI suggests that a formal set of rules must complement medical training and common sense for clinicians to effectively interpret the results of clinical research. Finally, EBI places a lower value on authority than the traditional paradigm of medical practice.

The evidence-based process involves a series of steps: (1) formulation of the clinical question, (2) identification of the medical literature, (3) critical appraisal of the literature, (4) summary or synthesis of the evidence, and (5) application of the evidence to derive an appropriate course of action.

An evidence-based practitioner must be able to understand the patients' circumstances or predicament (including issues such as their social supports and financial resources); identify knowledge gaps, and frame questions to fill those gaps; conduct an efficient literature search; critically appraise the research evidence; and apply that evidence to patient care.

The overall number of structured review articles in medicine has increased more than 40 times in the last two decades, according to a search of the publication terms “meta-analysis” (MeSH or tw) or “systematic review” (tw) in MEDLINE, derived from Ovid: from 3,255 articles published prior to 1994 to 22,302 articles up to 2004 to 122,232 articles.
up to 2015 (searched in late December 2015). However, of these review articles catalogued up to 2015, only approximately 4,470 (3.7%) have evaluated radiology-related topics, including conventional radiography, ultrasonography, computed tomography, magnetic resonance imaging, and radionuclide imaging. The proportion of systematic review/meta-analysis articles in radiology compared to the overall number in the entire field of medicine has slowly increased over the last two decades (1.6% in 1994 and 2.6% in 2004). However, these numbers still indicate a paucity of articles that summarize the best estimates of procedures’ effects, imaging as outcome measures in clinical effectiveness studies, diagnostic tests, or economic evaluations in radiology. This could be secondary to an insufficient body of primary evidence that can be reviewed for some topics in radiology, and/or relate to the fact that some reviews that contain a substantial number of low-quality primary studies may provide contradictory evidence on the effectiveness of interventions or accuracy of diagnostic tests.

### Definitions and Types of Reviews

Reviews are essential tools for researchers and clinicians who want to keep up with the evidence that has been accumulated in their fields. They can be unstructured (narrative reviews or commentary) or structured (systematic reviews, meta-analysis, pooled analysis). The latter type of reviews enables assessment of existing evidence on a topic of interest and concludes to support a practice, refute a practice, or identify areas for which additional studies are needed.

The most generic types of review articles are narrative overviews, followed by systematic reviews, meta-analyses, and pooled analyses. A narrative overview is a potentially biased nonstructured literature review on a specific topic that raises a broad research question. It provides a qualitative summary of the literature in the field. The Radiology series “State of the Art” and “How I Do It” are typical examples of narrative reviews where experts are invited to write an article for the journal.

A systematic review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, critically appraise relevant research, collect, and analyze data from studies included in the review. They are classified as “secondary” literature and should be distinguished from original published journal articles, which are classified as the “primary” literature. Table 7.1 compares the characteristics of narrative and systematic reviews.

Systematic reviews aim at estimating summary effects (synthetic goal) from a qualitative perspective and may or may not include a meta-analysis. The term meta-analysis is used when an attempt is made to estimate summary effects (synthetic goal) and differences (analytic goal) from a quantitative perspective by applying statistical methods. A pooled analysis is a meta-analysis based on individual-level patient or study data. Although pooling results of multiple studies reduces random error and increases the applicability of results across a broad range of patients, it risks violating the initial assumption of the analysis, which is to provide a nonbiased single best estimate of a patient’s prognosis, the effect of a treatment or diagnostic procedure, or the accuracy of a diagnostic test. The solution to this dilemma is to evaluate the extent to which results differ from study to study; namely, the heterogeneity of study results which is further discussed in the Chapter 16.

The art of conducting and evaluating a systematic review or meta-analysis requires previous knowledge on evidence-based concepts.

### Relevance

Unsystematic observations of clinicians constitute one source of evidence, and physiologic experiments another. Unsystematic clinical observations are limited by small sample size and by limitations in human processes of making inferences. Predictions...
about intervention effects on clinically important outcomes from physiologic experiments are usually right, but can be wrong. Observational studies are inevitably limited by the possibility that apparent differences in treatment effect are really due to differences in patients’ prognosis in the treatment and control groups.3 Given the limitations of unsystematic clinical observations and physiologic rationale, EBI is desirable.

The overall goal of a systematic review or meta-analysis is to combine results of previous studies to arrive at summary conclusions about a body of research.16 A properly conducted systematic review/meta-analysis can summarize large amounts of data. For health care providers, consumers, and policy makers who are interested in the bottom line of evidence, systematic reviews can help outline conflicting results of research. In radiology, systematic reviews or meta-analyses can be used to provide a summary estimate of effect size of a treatment that used imaging data to assess outcomes in observational or randomized controlled clinical trials, to estimate the clinical effectiveness of an imaging-guided therapy procedure, to synthesize results of economic evaluations that have used imaging data, or to evaluate the summary diagnostic accuracy of an imaging test. With regard to the latter purpose, clinicians, policy makers, and patients would like to know if the application of the test improves the outcome, what test to use or to recommend in practice guidelines, and how to interpret test results.17 Well-designed diagnostic accuracy studies can help in making these decisions, provided that they fully report their participants, tests, methods, and results.

Table 7.1 Comparison of characteristics of narrative and systematic reviews

<table>
<thead>
<tr>
<th></th>
<th>Narrative review</th>
<th>Systematic review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overarching research question(s) scope</td>
<td>Broad</td>
<td>Narrow (focused)</td>
</tr>
<tr>
<td>Authorship</td>
<td>Typically one or a small number of authors from a given discipline</td>
<td>Typically multiple authors from different disciplines that relate to the review scope</td>
</tr>
<tr>
<td>Source</td>
<td>Data provided by the authors, therefore often biased</td>
<td>Data obtained systematically to identify all relevant literature, therefore less prone to bias</td>
</tr>
<tr>
<td>Data appraisal</td>
<td>No objective assessment of the quality of the data reviewed</td>
<td>Objective assessment of the quality of the primary studies</td>
</tr>
<tr>
<td>Data synthesis</td>
<td>No statistical analysis of primary studies</td>
<td>If possible statistical analysis by meta-analysis; if heterogeneity of primary studies then meta-analysis is not possible and reasons for data heterogeneity should be explained</td>
</tr>
<tr>
<td>Data reliability</td>
<td>Replication of results not expected</td>
<td>Expected replication of results by using the presented methods in primary studies</td>
</tr>
<tr>
<td>Inferences</td>
<td>Conclusions reflect a small group of experts’ perspectives, therefore may be biased</td>
<td>Conclusions of overarching questions based on objective data analysis</td>
</tr>
</tbody>
</table>

Steps for Conducting a Systematic Review

Steps involved in a systematic review are similar to the phases of any other research undertaking formulation of the problem to be addressed, collection, critical appraisal (quality assessment), analysis of data from observational or randomized studies, and interpretation of results (assessment of heterogeneity, sensitivity, and subgroup analyses).
Protocol Phase
The initial step in the protocol phase is to define the main outcome of interest, such as, clinical effectiveness of diagnostic procedures or drugs in studies that use imaging as outcome measures, performance of diagnostic tests, or cost–benefit, cost-effectiveness, or cost-utility of treatment strategies or health care programs that involve diagnostic imaging tools. Before the start of the study, a detailed review protocol should be established that clearly states the question to be addressed, subgroups of interest, methods, and criteria to be used to identify and select relevant studies, and extract and analyze information. However, in some circumstances, unexpected or undesired results can be excluded by post hoc changes to the inclusion criteria, which should be documented in the review. Eligibility criteria for the review should define study participants, interventions, outcomes, study designs, and method quality of studies to be included in the review. An example on how authors can establish a systematic review/meta-analysis protocol to evaluate the diagnostic performance of ultrasonography (US) and computed tomography (CT) for the diagnosis of appendicitis in pediatric and adult populations18 is available in Table 7.2.

Review Phase
Identification of Studies
The search strategy for identifying relevant studies should be clearly defined considering multiple database sources: MEDLINE, EMBASE, EBM reviews, Cochrane Controlled Clinical Trials Register (CCTR), and bibliographic databases specific to such disciplines as nursing (CINAHL), behavioral sciences (PsycINFO), alternative medicine (MANTIS, AMED), physiotherapy (PeDRO), and oncology (CANCERLIT); checking of reference lists and personal files; hand searching of key journals; and personal communication with experts in the field. The search process should include MeSH terms pertaining to the population, intervention, comparison groups, and outcomes of interest as described in Chapter 10.

A comprehensive search should consist of the following steps:
1. Meta-analysis (pt)
2. Meta-anal: (textword)
3. Metanal: (textword)
4. Quantitative: review: OR quantitative: overview: (textword)
5. Systematic: review: OR systematic: overview: (textword)
6. Methodologic: review: OR methodologic: overview: (textword)
7. Review (pt) AND Medline (text word) [and other databases]
   1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7

It is highly recommended that authors prepare a flow diagram revealing the search and selection process used for the identification and quality assessment of articles (Fig. 7.1). Moreover, when feasible, investigators should use a “topic-only” search strategy and avoid restrictions with regard to articles written in certain languages as an attempt to prevent language bias.19,20

Selection of Studies
Decisions regarding the inclusion or exclusion of individual studies often involve some degree of subjectivity. It therefore is useful to have at least two readers checking eligibility of candidate studies, with disagreements resolved by consensus or by a third reviewer.

It is recommendable to keep a log of excluded studies with reasons for exclusions, which should be available on request from the authors of the review.

Assessment of “Risk of Bias” for Study Quality
Independent assessment of method quality of individual studies by more than one reader is recommended. Blinding of readers to investigators’ names and institutions, journal names, and acknowledgments is controversial because it is time consuming and potential benefits may not always justify the additional costs.21 The quality of primary studies can be measured with scales or checklists. Use of scales involves assigning each item a
Table 7.2 Example of a protocol (inclusion and exclusion criteria) for study selection in a systematic review/meta-analysis

<table>
<thead>
<tr>
<th>Study Feature</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>Segmentation of results according to age groups, with a maximum age of 20 years for children and young adults and a minimum age of 13 years for adults, and if this criterion was not fully met, the proportion of patients with outlying ages could not exceed 5% of the total sample size. Inclusion of both female and male patients (i.e., ratio of one sex to the other, &lt; 3:1)</td>
<td>Data for pregnant women</td>
</tr>
<tr>
<td>Target disorder</td>
<td>Appendicitis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Studies with a 15%-75% sample prevalence of appendicitis derivable from the reported results (i.e., true-positive plus false-negative results, divided by the total of true-positive, true-negative, false-positive, and false-negative results), as arbitrarily determined and checked by means of sensitivity analysis</td>
<td></td>
</tr>
<tr>
<td>Research design of primary studies</td>
<td>Prospective or retrospective studies evaluating the performance of abdominal ultrasound and/or CT</td>
<td>Case reports, case series, reviews, pictorial essays, unpublished data, abstracts, and letters to editor</td>
</tr>
<tr>
<td></td>
<td>Availability of data for the absolute number of true-positive, true-negative, false-positive, and false-negative findings either reported, derivable from the results, or communicated by the authors in response to our request</td>
<td>Focus on topics other than diagnostic test assessment, such as management decision issues or cost-effectiveness analyses</td>
</tr>
<tr>
<td></td>
<td>No language restriction</td>
<td></td>
</tr>
<tr>
<td>Prior tests</td>
<td>Exclude studies where patients had prior diagnosis of appendicitis (interval appendectomies)</td>
<td>When more than one study uses the same data or when the durations of studies overlap, the study with the larger sample size is selected to avoid duplication of data</td>
</tr>
<tr>
<td>Ultrasound test methods</td>
<td>Criteria for positive and negative test results defined. Imaging criteria for positivity for appendicitis that included visualization of an inflamed appendix (diameter &gt; 6 mm) noncompressible appendix at ultrasound, or, in the case of nonvisualization of the appendix, presence of inflammatory signs of appendicitis, such as an appendicolith, cecal thickening, arrowhead sign, or cecal bar (as seen on CT images)</td>
<td>Performance of more than one ultrasound examination per patient</td>
</tr>
</tbody>
</table>
### Table 7.2  Example of a protocol (inclusion and exclusion criteria) for study selection in a systematic review/meta-analysis (continued)

<table>
<thead>
<tr>
<th>Study Feature</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT test methods</td>
<td>Criteria for positive and negative test results defined. In studies evaluating the performance of CT scanning, a description of the technique used—namely, the use of oral, rectal, and intravenous contrast material with a limited or complete scan. Experience of operators described.</td>
<td>Performance of more than one CT examination per patient.</td>
</tr>
<tr>
<td>Reference test</td>
<td>Surgical/anatomopathologic or follow-up results. Criteria for positive and negative test results defined.</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 7.1** Hypothetical flow diagram revealing the search and selection process used for the identification and quality assessment of articles.

- 1000 search results identified in MEDLINE, EMBASE, DARE, and Cochrane Library
- 50 articles retrieved for full-text article review after duplicates were removed
- 38 full-text articles included in qualitative synthesis
- 20 articles included in quantitative synthesis (meta-analysis) if appropriate
- 12 articles excluded for the following reasons:
  - Duplicate articles (2)
  - Age of patients (4)
  - Too few patients (2)
  - Too few patients (3)
  - Not focused on disease (1)
- 23 search results identified through bibliographic references
numerical score; the sum of scores of these items then determines the overall quality of the study.\textsuperscript{22} Use of checklists involves scoring items as “yes” or “no” and assigning one point for each “yes” item; the final score for a study corresponds to the sum of these “yes” items.\textsuperscript{23} Theoretical considerations\textsuperscript{24,25} suggest that scales generally not be used to assess the quality of trials in meta-analyses, and sample checklists should be preferable. A grading display of the strength of evidence level according to type of design of the studies is shown in Fig. 7.2. In general, the type of design has an effect on the overall quality of the primary study. Double-blind, randomized, controlled, clinical trials provide the strongest evidence for causality relationship. Conversely, indirect evidence shown in case reports, expert opinion, and consensus committees provides the weakest evidence.\textsuperscript{26} Some potential sources of bias in primary studies include selection bias (caused by incomplete randomization or allocation of patients in the alternative and standard care groups), performance bias (caused by differences in care provided to patients exposed or not to an intervention or diagnostic procedure), detection bias (caused by differences in outcome assessment between two groups of patients), and attrition bias (caused by differences in withdrawal or participation rates of patients in randomized controlled trials).\textsuperscript{27}

Let us assume a hypothetical example of a clinical trial that compared the effectiveness of an imaging-guided therapy procedure (alternative arm of the study) and an open laparotomy procedure (standard care study arm) in terms of rates of postprocedure complications in subsets of patients with clinically suspected perforated appendicitis. Without a strategy to avoid biases, more patients with severe clinical symptoms were examined by means of CT than sonography before their surgical procedures in each of

![Fig. 7.2 Quality of primary studies in a meta-analysis according to the strength in causality related to study designs.](Image)
the study arms (selection bias). In addition, their imaging guided therapy and open laparotomy procedures were performed by both a radiology fellow and a staff radiologist, rather than by the on-call fellow only (performance bias), and their rate of complications in each procedure arm was evaluated by two observers, rather than by a single observer, as would be the standard approach for patients with less severe abdominal pain (detection bias). Finally, a greater proportion of patients with severe abdominal pain in both study arms had refused to participate in the study; therefore, some included patients may not have the outcome (attrition bias). In this example, despite the presence of biases in both arms of the study, results in patients with severe abdominal pain were systematically different from results in patients with less severe symptoms. These biases could have influenced the summary estimate of the effect of the procedures in subsets of patients with clinically suspected perforated appendicitis in a meta-analysis that included this particular study.

**Extraction of data:** A standardized data record form is needed for this purpose. More than one reader should extract the data to avoid or minimize errors and inadequate indexing of existing reports of studies recorded by different observers. An arbitrator should be required to reach agreement. The rate of disagreement between readers during the data-extraction process should be shown in the final report. To facilitate extraction and subsequent analysis, data should be noted in a study table designed with the research questions in mind (example in Supplementary Table 1). Important domains and elements to rate methodologic quality of individual studies in data synthesis of studies of clinical or technology implementation effectiveness and of diagnostic tests and economic evaluations should include:

- Study question
- Study population
- Randomization
- Blinding
- Intervention
- Outcomes
- Statistical analysis
- Results
- Discussion
- Funding (if appropriate)

### Data Synthesis of Studies of Clinical or Technology Implementation Effectiveness

For data synthesis on randomized controlled clinical trials as primary studies, checklists of risk of bias (The Cochrane Collaboration’s tool for assessing risk of bias) can assess the quality of primary studies. After studies have been selected and critically appraised and data have been extracted, characteristics of included studies and individual results should be expressed in a standardized format to allow for comparison between studies. If the outcome is binary (e.g., disease vs. no disease; intervention vs. standard practice procedure), odds ratios, relative risks, or risk differences can be calculated. If the outcome is continuous (e.g., percentual enhancement of a tissue after contrast administration), mean difference, standardized mean difference, or correlation coefficients can be applied. Odds ratios have convenient mathematical properties because they do not have inherent range limitations associated with high baseline rates and are suitable for statistical manipulation as the antilog of coefficients. Details on this are available in Chapter 16. Nevertheless, relative risks usually are preferred over odds ratios because they are more intuitively understandable.

Before pooling results of individual studies using an effect measure (e.g., odds ratio or relative risk), the investigator should evaluate for the presence of heterogeneity within and between studies.

### Data Synthesis of Studies of Diagnostic Tests

Methodologic quality assessment of individual studies in systematic reviews is therefore necessary to identify potential sources of bias and to limit the effects of these biases.
on the estimates and the conclusions of the review. Methodologic quality of a study has been defined as “the extent to which all aspects of a study’s design and conduct can be shown to protect against systematic bias, non-systematic bias that may arise in poorly performed studies, and inferential error.”

The Standards for Reporting of Diagnostic Accuracy (STARD) checklist was not developed as a tool to assess the quality of diagnostic studies. This 25-item checklist has been used to evaluate the quality of reporting of diagnostic studies by ensuring that all relevant information is present.

However, many items in the checklist are included in recently developed tools for quality assessment of diagnostic accuracy (the Quality Assessment of Diagnostic Accuracy Studies [QUADAS-2] tool) and reliability (the Quality Appraisal of Reliability Studies [QAREL] tool, Supplementary Table 3). The QUADAS-2 tool is structured as a list of two domains (risk of bias and applicability) and 14 questions on diagnostic accuracy that should each be answered “yes,” “no,” or “unclear.” Under both domains items cover patient selection, index test, reference standard, verification and review bias, clinical review bias, incorporation bias, test execution, study withdrawals, and intermediate results. Additionally, the risk of bias domain also covers flow/timing. The QAREL tool includes 11 items that explore seven principles. Items cover the spectrum of subjects, spectrum of examiners, examiner blinding, order effects of examination, suitability of the time interval among repeated measurements, appropriate test application and interpretation, and appropriate statistical analysis.

The results of quality appraisal can be summarized to offer a general impression of the validity of the available evidence. Review authors should not use an overall quality score, as different shortcomings may generate different magnitudes of bias, even in opposing directions, making it very hard to attach sensible weights to each quality item. A way to summarize the quality assessment is shown in Fig. 7.3, where stacked bars are used for each QUADAS-2 item. Another way of presenting the quality assessment results is by tabulating the results of the individual QUADAS-2 items for each single study. The effects of the STARD guidelines for complete and transparent reporting are only gradually becoming visible in the literature.

Fig. 7.3 Graphical display of QUADAS-2 assessment of methodological quality. The majority of the “reference standard” domains were rated as “not applicable” as these studies were not identified as diagnostic accuracy studies. The “patient selection” item of the “risk of bias” domain was rated as “unclear” for the majority of studies as the sampling method was often not reported. However, due to the low prevalence of hemophilia, the authors of this review thought that it would often not be feasible to carry out a randomized sampling process. Thus, we primarily used the “index test” and the “flow and timing” items to determine risk of bias.
Data Synthesis of Economic Evaluations

The aim of data synthesis of economic evaluations is to summarize the evidence about the efficiency of health care provision to reduce the uncertainty about relative benefits and costs associated with alternative interventions. Different forms of analysis can be applied to economic evaluations that involve diagnostic imaging modalities: cost–benefit analysis, cost-effectiveness analysis, and cost–utility analysis. Cost–benefit analysis measures loss to society or net gain of a new program and thus considers allocative efficiency. Cost-effectiveness analysis assesses technical efficiency and compares alternative approaches to care. Cost–utility takes into consideration the utility of health gain. All forms of economic analysis involve summing up of information about quantity and value of inputs and outcomes experienced by those undergoing the alternative forms of health care.

Methods for synthesis of economic evaluations are not as advanced as those used for clinical effectiveness. Nevertheless, decision models (Chapter 8) are frequently used to conduct economic evaluations alongside systematic reviews, as only rarely is all the evidence required for an economic evaluation available from one source.

The first step is to evaluate the quality of evidence about clinical effects. Several tools for assessment of the quality of economic evaluations are currently available, including guidelines for authors and peer reviewers of economic submissions to the British Medical Journal (BMJ checklist) and the Quality of Health Economics Studies (QHES) checklist.

The Campbell & Cochrane Economics Methods Group recommends a two-stage approach to assessment of health economic evaluations, utilizing the Cochrane Collaboration’s tool for assessing risk of bias (stage one) and, depending on the approach to economic evaluation, the BMJ checklist and/or the criteria list for assessment of methodological quality of economic evaluations or the quality assessment in decision-analytic models (Philips checklist) to inform each stage.

The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) system provides a comprehensive framework for rating the quality of evidence on both health and other effects, collected from randomized controlled trials and observational studies. It can also be used to address evidence on resource use and costs, drawn from previously published economic evaluations. Supplementary Table 4 shows the GRADE criteria for rating quality of overall (a) and specific economic (b) evidence. If reliable evidence of equivalence or increased effectiveness is present, the next step is to assess the available economic evidence. The initial analysis of economic evidence usually is nonquantitative and similar to the analysis pursued for studies of clinical effectiveness or diagnostic tests. Although the absolute level of resource utilization and costs may differ among countries in which individual studies were conducted, the analysis should focus on whether the effect of the treatment or diagnostic procedure differs among countries. Results can be tabulated under several headings, including quality of effectiveness evidence (experimental, observational, mixed, and so on), magnitude of effectiveness (benefits), sources of cost data (hospital/patient records, reimbursement tariffs, literature review, and so on), and costs (direct and indirect). In addressing the question of whether to pool estimates of treatment or diagnostic procedure effects over studies conducted in different countries, we need to evaluate the extent to which this effect is common among studies included in the meta-analysis and assess whether no substantial evidence suggests heterogeneity. To facilitate comparison between countries and time periods, it may be possible to standardize cost data by inflating or deflating cost to one specific year if sufficient details have been reported in the primary studies.

Cochrane Collaboration

The Cochrane Collaboration is an international nonprofit organization of more than 31,000 dedicated contributors from over 120 countries that prepares, maintains, and disseminates reviews of health care interventions that are evidence based, easily accessible,
internationally developed, quality controlled, clinically useful, and periodically updated. They decided in 2003 to make preparations for including systematic reviews of diagnostic test accuracy in their Cochrane Database of Systematic Reviews (CDSR). Currently, this organization focuses on specific groups of disorders or systems or disciplines (e.g., musculoskeletal, neonatal, neuromuscular disease, public health, sexually transmitted infections, etc.) across the world. These centers help organize and register review groups, facilitate collaboration among reviewers, provide training and consultation, establish liaisons, and promote Cochrane collaboration.

- **Users’ Guides to the Medical Literature**

Evidence-based technical skills and humane perspectives will lead physicians to become effective advocates for their patients both in the direct context of the health system in which they work and in broader health policy issues. This advocacy may involve changing the system to facilitate evidence-based practice; for example, by improving infrastructure for access to high-quality information to guide clinicians at the bedside. A continuing challenge for EBI and for medicine in general will be to better integrate the new science of clinical medicine with the time-honored craft of caring for the sick.

The biggest obstacle to evidence-based practice remains time limitation. One way to go around this obstacle is to assess evidence by systems, using summaries that link a number of synopses related to the care of a particular patient problem (arthritis, stroke, etc.) or type of patient.

The Users’ Guides to the Medical Literature provide clinicians with the tools to distinguish stronger from weaker evidence, stronger from weaker syntheses, and stronger from weaker recommendations for moving from evidence to action. Much of the Guides are devoted to helping clinicians understand study results and enumerate the benefits, side effects, toxicity, inconvenience, and costs of treatment options, both for patients in general and for individual patients under their care. A clear understanding of the principles underlying evidence-based practice should aid clinicians in applying the Users’ Guides to facilitate their patient care.

- **Interpretation of Results: Clinical Practice Guideline Development**

Evidence-based approaches can inform health policy-making, day-to-day decisions in public health, and systems-level decisions such as those facing managers at the hospital level. Clinical practice guidelines have been developed to improve the process and the outcome of health care and to optimize resource utilization. They are based on the interpretation of the results offered in the systematic review and should help readers to understand the implications for practice. By addressing issues such as therapy, harm, prognosis, and diagnosis, they can aid in health care decision making at many levels. According to Li et al, the sequential process for developing guidelines includes:

1. Assembling a multidisciplinary team of experts.
2. Establishing and agreeing on the process: deriving and refining clinical questions and current clinical practice.
3. Identifying and updating existing systematic reviews.
4. Preparing a systematic review of the best available evidence for important outcomes of the clinical problem: for each subpopulation or risk group, based on the results of systematic reviews, including a summary of findings.
5. Appraising methodological quality of evidence of available literature.
7. Balancing net health benefits and costs: Are incremental health benefits worth the costs?
8. Identifying evidence gaps and prioritizing comparative effectiveness research.
10. Implementing and evaluating strategies that address barriers to change, and linking guidelines to form clinical pathways.
11. Disseminating guidelines and clinical pathways.

**Critical Appraisal of Systematic Reviews**

The inconsistent quality of systematic reviews that are published in the literature has been previously documented and several guidelines and checklists have been published in an attempt to improve both the quality and completeness of reporting of systematic reviews and meta-analyses.

Guidelines include specific elements designed to improve the completeness for reporting systematic reviews and meta-analyses. The first major guideline for systematic review and meta-analysis reporting was the Quality of Reporting of Meta-Analyses (QUOROM), which was published in 1996 and this was recently updated by the publication of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement in 2009. The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram (Supplementary Table 5). Similarly, multiple tools have been proposed to assess the quality of systematic reviews and meta-analyses. The Assessing the Methodological Quality of Systematic Reviews (AMSTAR) tool, published in 2007, is a 37-item assessment tool that was formed by combining (1) the enhanced Overview Quality Assessment Questionnaire (OQAQ), (2) a checklist created by Sacks, and (3) three additional items recently judged to be of methodological importance. It was previously validated as a tool to assess the quality of published systematic reviews. A recent study has shown a strong positive correlation ($r = 0.86$) between the PRISMA (Preferred reporting quality of systematic reviews and meta-analyses) and AMSTAR (methodologic quality of systematic reviews) results. Based on final assessments of quality of available primary studies and/or systematic reviews on a given topic, one can use grading systems of recommendations such as those suggested by the Canadian Task Force on Preventive Health Care which currently uses Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidelines (Supplementary Table 7), U.S. Preventive Services Task Force, among others.

**Bias**

These biases relate to systematic errors in primary studies. Publication bias is a systematic error in which the likelihood of publication of a study is influenced by the significance of its results. It occurs because studies with statistically significant results are more likely to be published than studies with nonsignificant findings. Therefore, results from unpublished studies may systematically differ from those of published trials. For this reason the search of unpublished studies and abstracts of scientific meetings and the inclusion of articles written in different languages are highly encouraged in the conduct of systematic reviews. Additional biases in systematic reviews include biases in selection of studies, in interpretation of “risk of bias” among included studies, and in decisions on pooling results of individual studies.

Investigators can deal with potential biases in the data used in primary studies by conducting sensitivity analysis which can assess the impact of the quality of the evidence included in the systematic review on its results. Should this analysis reveal that the results of the model are sensitive to different qualities of evidence, some solutions are proposed.

1. Evidence obtained from studies that fall below a prespecified quality threshold can be excluded.
2. Weights can be given to studies according to their quality.
3. Random effects modeling of bias can be employed.
4. Full bias modeling can be employed in an attempt to identify all sources of potential bias in the available evidence, obtain external information on the
likely form of each bias, and construct a model to correct the data analysis accordingly.

Limitations

1. Systematic reviews are limited by the quality and availability of primary test accuracy studies that address relevant questions. More studies that recruit suitable spectrums of participants, make direct comparisons between tests, use rigorous methodology, and clearly report their methods and findings are needed. Systematic reviews on the same topic have been known to reach different conclusions.

2. More development is needed in the area of interpretation and presentation of the results of diagnostic test accuracy reviews. We should also explore how well the concept of diagnostic accuracy applies to other forms of testing, such as prognosis, prediction, and monitoring, and to new test modalities, such as microarrays and genotyping. Policy makers and guideline developers may be interested in the comparative accuracy only, as well as in additional information, such as the costs and burden of testing, or in new test modalities.

3. Publication bias: articles in which results are positive are more likely to be published in English, to be published rapidly, and to be cited.

4. Studies of diagnostic tests are generally less methodologically rigorous than are studies of treatment effects in randomized controlled trials, which makes the combination of studies of diagnostic tests problematic. Methods for studies of diagnostic tests still lag behind those employed in randomized trials.

5. The poor quality of some trials means that we must disregard their findings and not attempt to pool them in a meta-analysis.

6. Statistical methods for meta-analysis of measures of diagnostic accuracy are less developed than those for assessment of treatment effects.

Knowledge Translation of Research Findings

We will discuss the methods available for effective translation and dissemination of evidence-based information at the individual patient, health care system, health policymaking, and stakeholder levels.

Once summative outcomes of specific topics in health sciences are accrued by means of systematic reviews and meta-analyses, the next natural step is the translation of these outcomes into meaningful patient care outcomes across multiple contexts. Estimates indicate that two-thirds of organizations’ efforts to implement change fail. As a result of these evidence-practice and policy gaps, patients fail to benefit optimally from advances in health care and are exposed to unnecessary risks of iatrogenic harms, and health care systems are exposed to unnecessary expenditure resulting in significant opportunity costs that result in poorer quality of life and loss of productivity both personally and at the societal level.

Over the last 10 to 15 years, there has been increasing international policy and research attention on how to reduce the evidence-practice and policy gap. Across different health care systems, different terms describe these efforts including quality assurance, quality improvement, knowledge translation, knowledge utilization, knowledge transfer and exchange, innovation diffusion, implementation research, research utilization, evidence-informed policy, and evidence-informed health systems.

Findings from clinical and health services research cannot change population health outcomes unless health care systems, organizations, and professionals adopt them in practice. Knowledge translation (KT) is about raising knowledge users’ awareness of research findings and facilitating the use of those findings. Nevertheless, without the utilization of implementation research strategies many interventions found to be effective in health services research studies fail to translate into clinical practice.

The term translational research relates to the T1 bench to bedside process of transferring basic science knowledge into new drugs.
and technologies. T2 translational research is the process of taking current scientific knowledge and ensuring it is applied in routine (clinical) care.92

Implementation research is the scientific study of methods to promote the systematic uptake of clinical research findings and other evidence-based practices into routine practice, and hence to improve the quality (effectiveness, reliability, safety, appropriateness, equity, efficiency) of health care or well-being. It includes the study of influences on health care professional and organizational behavior.93 Moving research into practice/policy is only feasible if strength of evidence is sufficient.

Barriers to implementation may arise at multiple levels of health care delivery: the patient level, the provider team or group level, the organizational level, or the market/policy level.94 For this reason health services researchers are increasingly recognizing the critical role of implementation science, the so-called “knowledge translation science.”95 Currently, many grant agencies consider “knowledge translation (KT)” as an essential section of a research proposal.

Knowledge translation is conducted by:

- Deriving knowledge from primary studies, such as randomized clinical trials (knowledge inquiry).
- Synthesizing primary studies to form secondary knowledge (systematic reviews, meta-analysis, scoping reviews).
- Generating knowledge tools or products (third generation knowledge) such as practice guidelines, decision aids, multi-language translation of imaging protocols and scoring systems, production of videos, or care pathways based on best available evidence distilled from synthesized knowledge.

Two broad categories of KT are recognized: integrated KT (iKT) and end-of-grant KT.96 Most funding opportunities are built on the conceptual definition of KT, including one or both categories of KT. In iKT, potential knowledge users are engaged throughout the research process. This approach should produce research findings that are more likely to be directly relevant to and used by knowledge users. It should also incorporate a dissemination plan to share the results of the project with other interested knowledge users. Patient engagement is increasingly recognized as a core driver toward a high performance health system and should be included in iKT. With end-of-grant KT, the researcher develops and implements a plan for making potential knowledge-user audiences aware of the knowledge that is gained during a project. End-of-grant KT can involve more intensive dissemination activities that tailor the message and medium to a specific audience and, even further along the spectrum, can involve moving research into practice in cases where the strength of the evidence is sufficient. End-of-grant KT covers any activity aimed at diffusing, disseminating, or applying the results of a research project, as defined below. Diffusion relates to “let it happen” through publications and presentations in scientific meetings. Dissemination related to “help it happen” by tailoring the message and medium to an audience. Appropriate methods to translating research findings range from simple communication activities (diffusion, dissemination) to more intensive knowledge application efforts, such as workshops, academic detailing, and tool development. Among traditional tools for end-of-grant KT are conference presentations and publications in peer-reviewed journals which have often been the principal modes of communication to researchers. These forms of KT remain the best approach for research at the early stages of discovery, when the knowledge has more relevance to academics who are contributing to a body of evidence that is not yet appropriate for application. Publishing in open-access journals or repositories has the potential of reaching a much broader audience, thus increasing the likelihood of research uptake by those in the academic community as well as knowledge users and the general public. When there are potential knowledge-user audiences beyond the research community, end-of-grant KT activities should be more intensive and emphasize nonacademic modes of communication: the language of
publications should be adapted to the target audience (e.g., lay language) and can be presented in popular formats, such as websites or creative media (e.g., film, theatre, infographics). Sharing of knowledge may be done face to face in a meeting/workshop setting by a knowledge broker (an individual specializing in the communication of findings to knowledge users, in their context) or via emerging online technologies. Long-form online social media platforms include blogs, YouTube, and podcasts, and short-form platforms include Twitter, Facebook, microblogging. To disseminate more broadly to the general public, media such as television, radio, and print may be engaged. The development of products and services based on research results (e.g., tutorial videos from consensus meetings), including commercialization activities, is also a form of end-of-grant KT.

Lavis and colleagues have identified five key questions that can be used as a guideline for knowledge translation activities:

1. What should be transferred?
2. To whom should research knowledge be transferred?
3. By whom should research knowledge be transferred?
4. How should research knowledge be transferred?
5. With what effect should research knowledge be transferred?

What should be transferred? This question challenges knowledge translators to identify the key messages for different target audiences and to fashion these in language and knowledge translation products that are easily assimilated by different audiences. Over the past decade, a variety of different products have been developed targeting different audiences (e.g., decision aids for patients, clinical practice guidelines for health care professionals, and actionable messages and policy briefs for policy makers).

To whom should research knowledge be transferred? The relative importance of knowledge translation to different target audiences will vary by the type of research being translated. For example, primary target audiences for knowledge translation of the results of basic science include other researchers and industry; whereas primary target audiences for knowledge translation of the results of population health research include other researchers, administrators, and policy makers. The relative importance of different target audiences will also vary by the results of the research. For example, the primary target audiences for clinical research demonstrating lack of benefit or harms from an imaging-guided therapy device sufficient to warrant its withdrawal might be national policy makers (including regulatory bodies) and industry. Whereas, the primary target audiences for clinical research demonstrating benefits from the use of one intervention device over another one to suggest its widespread use might be patients, health care practitioners, local administrators, as well as national policy makers and industry.

By whom should research knowledge be transferred? The messenger in knowledge translation efforts may be an individual (e.g., health care practitioner, researcher, or consumer), group, organization, or even health care system. The most appropriate messenger will vary according to the target audience and research knowledge being transferred. An authoritative endorsement by a respected physician organization or physician colleague is shown to influence physicians’ use of clinical practice guidelines in practice.

How should research knowledge be transferred?

Planning for Knowledge Translation

Most knowledge translation models suggest that planned knowledge translation is more likely to be successful if an assessment of the likely barriers and facilitators informs the choice of knowledge translation strategy.

Identifying Barriers to Knowledge Translation

Common barriers across target groups include issues relating to knowledge management, such as the sheer volume of research evidence currently produced, access to research evidence sources, time to read evidence sources, and skills to appraise and understand research
There are diverse methods for identifying potential barriers including qualitative approaches (individual interviews, focus groups), surveys, and direct observation. However, there are no standard approaches available yet. As a result, those involved with knowledge translation activities need to use their judgment about how best to elicit barriers given their understanding of the context and potential barriers and resources available to them.

Choosing Interventions

Individuals involved in knowledge translation need to: identify modifiable and nonmodifiable barriers relating to behavior; identify potential adopters and practice environments; and prioritize which barriers to target based upon consideration of mission critical barriers.

Printed Educational Materials

This activity is defined as the “distribution of published or printed recommendations for clinical care, including clinical practice guidelines, audio-visual materials and electronic publications. The materials may have been delivered personally or through mass mailings.”

Educational Meetings

This activity is defined as the “participation of health care providers in conferences, lectures, workshops or traineeships.” An important distinction is between didactic meetings (that largely target knowledge barriers at the individual health care professional/peer group level) and interactive workshops (that can target knowledge, attitudes, and skills at the individual health care professional/peer group level). Educational meetings are commonly used, with the main cost related to the release time for health care professionals, and are generally feasible in most settings.

Educational Outreach or Academic Detailing

These activities are defined as the “use of a trained person who meets with providers in their practice settings to give information with the intent of changing the providers’ practice. The information given may have included feedback on the performance of the provider(s).”

Local Opinion Leaders

This activity is defined as the “use of providers nominated by their colleagues as ‘educationally influential.’” The investigators must have explicitly stated that their colleagues identified the opinion leaders. Opinion leadership is the degree to which an individual is able to influence other individuals’ attitudes or overt behavior informally in a desired way with relative frequency. This informal leadership is not a function of the individual’s formal position or status in the system; it is earned and maintained by the individual’s technical competence, social accessibility, and conformity to the systems’ norms.

Audit and Feedback

Audit and feedback are “any summaries of clinical performance of health care over a specified period of time” to change health professional behavior, as indexed by “objectively measured professional practice in a health care setting or health care outcomes.” The summary may also have included recommendations for clinical action. The information may have been obtained from medical records, computerized databases, or observations from patients. The subsequent feedback of and resulting action planning based on the audit summary are also important elements of an audit and feedback intervention.

Reminders

Reminders are patient or encounter specific information, provided verbally, on paper or on a computer screen, which is designed or intended to prompt a health professional to recall information. This would usually be encountered through their general education, in the medical records or through interactions with peers, and so remind them to perform or avoid some action to aid individual patient care. Computer-aided decision support and drugs dosage are included.
With what effect should research knowledge be transferred?

Knowledge translation may vary across different stakeholder groups. For example, knowledge translation targeting policy makers and consumers should ensure that consideration of research evidence is a key component of their decision making, but recognize that there are other legitimate factors (e.g., the policy context for policy makers, values, and preferences of individual patients) that need to be considered. Thus, the resulting decision is likely to be evidence-informed but may not be particularly evidence-based. However, knowledge translation targeting professionals should result in practice that is more evidence-based and should be reflected in changes in professional behaviors.

■ Highlights of Key Points

1. Evidence-based reviews are classified as the “secondary” literature and should be distinguished from original published journal articles, which are classified as the “primary” literature.
2. “Narrative review” is a descriptive overview by an expert (or experts) of a selection of studies with published findings. Systematic review is an article that summarizes other articles (known as the “primary studies”), extracting relevant information from them (about methods and findings), and summarizing their results. The difference between a properly conducted systematic review and a narrative review is that there is a formal, careful search procedure for the former so that all relevant research is identified.
3. Systematic reviews and meta-analyses can be useful decision-making tools for health care providers, consumers, and policy makers by objectively summarizing large amounts of information, identifying gaps in medical research, and identifying beneficial vs. harmful, effective vs. ineffective, accurate vs. inaccurate interventions and diagnostic tests.
4. There is a step-by-step process for systematic review of a diagnostic test.
5. The Cochrane collaboration is “an international organization that aims to help people make well-informed decisions about health care by preparing, maintaining, and promoting the accessibility of systematic reviews of the effects of health care interventions.”
6. Finding the answers to questions that are relevant to policy makers and consumers requires effective collaboration among medical researchers and other decision makers.
7. Strategies that increase consumer and policy maker awareness on identification of systematic reviews, that allow critical appraisal of reviews, and that incorporate reviews into the decision-making process should be further developed.
8. Scarce published research has evaluated the quality of available systematic reviews and the direct impact of systematic reviews on the health care decisions made by policy makers and consumers. “There is an important distinction between the best evidence in theory and the best available evidence in practice.”
9. Concept of KT/implementation science: the study of how to promote uptake of research/knowledge in decision making.
10. Concept of integrated KT research: research approaches that engage potential knowledge users as partners in the research process. Requires a collaborative or participatory approach to research that is action oriented and is solutions- and impact-focused.
11. Concept of end of project KT: the researcher develops and implements a plan for making knowledge users aware of the knowledge generated through a research project.

Tips for Undertaking a Systematic Review

1. Develop a protocol (explicit plan) for your review.
2. Conduct an effective database search strategy with the help of an experienced librarian.
3. If the number of available primary studies is too extensive consider refocusing the overarching questions of the review to make the review time manageable.
4. Keep an ongoing system of quality assurance (with experienced investigators both in the field of radiology the review is about and in systematic review methodology) throughout the process.
5. Develop an extraction datasheet that can reliably record the key information from included studies prior to the commencement of the review including explanations for exclusion of candidate primary studies.

**Tips for Conducting Knowledge Translation**

1. Identify a problem, and review and select knowledge related to this problem.
2. Adapt knowledge to local context.
3. Assess barriers and facilitators to knowledge use.
4. Select, tailor, and implement intervention to address barriers to knowledge use.
5. Monitor knowledge use.
6. Evaluate outcomes of knowledge use.
7. Develop mechanisms to sustain knowledge use.

### Historical Moments

The logo of the Cochrane Collaboration illustrates a forest plot of results of seven randomized controlled trials on the effects of administering steroids to women who were about to give birth prematurely. The first trial was reported in 1972 and the last in 1980, and the point estimate derived from these shows the clear benefit of giving steroids had these trials been reviewed systematically at that time.

Unfortunately, a systematic review was not performed until a decade later, and thousands of babies died despite the availability of information that could have saved them.

One of the first systematic reviews of controlled clinical trials in 1983 demonstrated the disastrous effects of administering class I antiarrhythmic agents after myocardial infarction despite the attractive theoretical reasons for administration.116,117

**References**

10. Staunton M, Dodd JD, McCormick PA, Malone DE. Finding evidence-based answers to practical questions in radiology: which patients with inoperable hepatocellular carcinoma will survive longer following transcatheter arterial chemoembolization? Radiology 2005;237:404–413


42. Marsh K. The role of review and synthesis methods in decision models. In: Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C. Evidence-based...


53. Guyatt GH, Sackett DL, Cook DJ for the Evidence-based Medicine Working Group. Users’ guides to the medical literature. II - How to use an article about therapy or prevention. Part B. What were the results and will they help me in caring for my patients? JAMA 1994;271:59–63


55. Richardson WS, Detrsky AS, for the Evidence-based Medicine Working Group. Users’ guides to the medical literature. VII. How to use a clinical decision analysis. Part B. What are the results and will they help me in caring for my patients? JAMA 1995;273:1610–1613


57. O’Brien BJ, Heyland DK, Richardson WS, Levine M, Drummond MF, for the Evidence-Based Medicine Working Group. Users’ Guides to the Medical Literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? JAMA 1997:277:1802–1806


60. McAlister FA, Straus SE, Guyatt GH, Haynes RB. Users’ guides to the medical literature: XX. Integrating research evidence with the care of the individual patient. Evidence-Based Medicine Working Group. JAMA 2000;283:2829–2836


62. Drummond MF, Richardson WS, O’Brien B, Levine M, Heyland DK, for the Evidence-Based Medicine Working Group. Users’ Guides to the Medical Literature. XII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? JAMA 1997;277:1552–1557


83. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. CMAJ 1997;156:1411–1416


85. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274:645–651


97. Lavis JN, Robertson D, Woodside JM, McLeod CB, Abelson J. How can research organizations more effectively transfer research knowledge to decision makers? Milbank Q 2003;81:221–222
99. Lavis JN, Robertson D, Woodside JM, McLeod CB, Abelson J. How can research organizations more effectively transfer research knowledge to decision makers? Milbank Q 2003;81:221–222
**Learning Objectives**

- To understand and be able to describe the limitations to making optimal clinical decisions.
- To know how to interpret a simple decision tree.
- To understand the meaning of expected value and be able to calculate this from a simple decision tree.

**What Makes Clinical Decisions Difficult?**

Clinical decision making is complex because critical, complex factors involving risk and prognosis must be weighed in order to choose the best option. Central to determination of risk and prognosis is the concept of probability. The probability of an event measures how likely it is to occur, expressed as a number between 0 and 1 equal to the proportion of possible outcomes in which the event of interest occurs. A probability can be subjective, quantifying an individual’s belief in the likelihood of an event, or based on actual frequency. In clinical practice, the probability of occurrence of an event for one patient is usually based on the frequency of the event in a group of similar patients. The challenges of thinking in probabilistic ways include choosing the correct group of patients with similar characteristics on which to make an inference about an individual patient.

A potentially large number of patient characteristics such as age, sex, co-morbid illnesses, physical findings, symptoms, disease history, and test results play an important role in determining precise assessments of risk and prognosis. For a radiologist, clinical decision making often concerns whether or not to image a patient, which imaging test to use, and making a diagnosis from the imaging findings. These decisions involve weighing the likelihood that the patient has the disease, the accuracy of the test for detecting the disease, potential harms that may result from the test, and the added value of the test in contributing to management. As a result, clinical decisions in radiology typically necessitate that health care providers
Decision Analysis

Medical practice. Clinicians absorb a vast amount of information during their early career training and new information related to clinical care is generated at a rapid pace. Furthermore, the clinical setting is characterized by a great degree of complexity. Important factors like the prevalence of disease vary widely for different populations as defined by age, sex, ethnicity, socioeconomic status, place of birth, and comorbid illness, and have an effect on the predictive values of diagnostic tests. Patients with the same disease may have different symptoms, and thus the need to remain vigilant due to the possibility of unique disease presentations increases complexity.

In one experiment, physicians from Australia and England were presented with clinical scenarios for chest pain, leg swelling, and atrial fibrillation. Each was presented with a vignette describing patient characteristics, symptoms, and clinical findings. Based on patient features, physicians were asked to estimate pre-test probability of diseases including deep venous thrombosis and stroke—conditions where imaging would be used to clarify the diagnosis. Estimates varied widely: for the clinical scenarios with the best performance, only approximately 55% of physicians were within 20% of the correct risk estimate. Results like these have important implications for radiologists, since an inaccurate estimation of pre-test probability for disease significantly impacts the predictive value of a diagnostic test. This further complicates the decision of whether or not a particular diagnostic test should be performed in a given clinical scenario by influencing the degree to which the test outcome adds value to patient management.

Difficulties in Risk Estimation

Individuals have difficulties accurately estimating risks. Individuals tend to overestimate the risk of events with a low probability and underestimate the risk of events with a high probability. For example, individuals may be more concerned about the risk of developing cancer from exposure to low doses of radiation and less concerned about the risk of cancer associated with obesity. This distortion in risk perception may lead someone to be more...
vigilant about advocating for increased government controls on radiation emissions near a nuclear power plant and less vigilant maintaining a healthy weight and eating nutritiously.

It is difficult for both clinicians and patients to think in probabilistic ways. For example, when doctors in Ontario were asked to estimate the 8-year probability of developing heart disease (angina, myocardial infarction, or coronary death) for hypothetical patients, they systematically overestimated each patient’s absolute risk.2 The physicians also overestimated the reduction in risk association with modifying one or more risk factors. Interestingly, both this and another study found that the physicians performed better when estimating the relative risk of disease due to the presence of an important risk factor than when estimating the baseline risk and the risk reductions for specific interventions.2,3 Only 24% of physicians had accurate risk estimates, but demonstrated greater accuracy in estimating the expected risk reduction from therapy, with 43% estimating relative risk accurately. The degree of overestimation was larger for low-risk scenarios, intermediate for medium-risk scenarios, and smaller for high-risk scenarios. A European study highlighted the difficulty faced by radiologists when attempting to quantify patients’ risk of developing contrast-induced nephropathy from exposure to contrast agents.4 The majority of radiologists surveyed used unstandardized and invalidated criteria for assessment of risk of contrast-induced nephropathy, which led to important inaccuracies in their risk estimates. Most radiologists underestimated the risk of contrast-induced nephropathy, and this may have significant clinical implications.

Uncertainty

The concept of uncertainty is distinct from risk, but equally important in clinical decision making. Han and colleagues define uncertainty as “the subjective perception of ignorance that is experienced by patients and health professionals in differing ways and degrees, motivates action, and elicits a variety of psychological responses.”5,6 In clinical decisions, there are two sources of uncertainty about probabilities or risks: (1) uncertainty about the exact value of a probability; and (2) uncertainty about what will happen in the future, even when there is certainty about the exact probability value. The latter type of uncertainty is easily recognizable. When tossing a coin, the probability of heads is known to be 0.5, but with any given toss the outcome is uncertain. Even if we accurately estimate a patient’s risk of developing cancer, unless the probability is 100%, we are still uncertain about what will actually happen to the patient. The former type of uncertainty, uncertainty about the exact probability value, arises largely from the complex factors involved in estimating risk and prognosis that have already been discussed. An honest assessment of the complexities involved in risk assessment and a thorough understanding of human cognitive limitations would lead to an appreciation of uncertainty.

Evidence shows that clinicians tend to be overly confident that their risk estimates are correct. In the United Kingdom, nursing students and experienced nurses were presented with 25 vignettes and were asked to estimate the risks of adverse events.7 The vignettes outlined vital signs and consciousness levels for hypothetical patients. Both groups performed poorly at estimating the risk of an adverse event. Experienced nurses had greater confidence in their judgments than students, but there were no statistical differences in the rates of incorrect risk estimates. Clinicians overly confident about risk estimates may provide suboptimal guidance to patients.

Patients may also be overconfident about their risk estimates. In diagnostic testing, patients’ preferences and preconceived notions about risks of imaging play an important role in decision making. A study performed in the pediatric setting highlighted the knowledge gap experienced by parents of children with head injuries regarding adverse events associated with radiation. Although approximately 50% of parents surveyed claimed to be aware of an association between CT imaging and potential cancer risk, over 90% of parents were initially willing to allow their children to undergo CT scanning of the brain; this number decreased to approximately 70% after discussion of potential long-term cancer risks.8
The challenges of conducting research to determine the exact values of clinical probabilities have been ignored in this chapter. The resources for financing research are scarce, and even for carefully designed studies, there are challenges in interpreting results and applying the results in one population to another population. The challenges of applying the tools of epidemiology to estimate risk and determine prognosis are addressed in earlier chapters of the textbook. This chapter is focused on a discussion of cognitive limitations in recalling or estimating probabilities and the possible way in which decision analysis can address these limitations.

Heuristics and Biases

The work of Amos Tversky and Daniel Kahneman demonstrates that when individuals make decisions, they frequently rely on intuitive reasoning in place of more deliberative processes in order to derive predictions and arrive at judgments faster. These intuitive short cuts are known as heuristics and they can lead to systematic errors in judgment that are called biases. Use of heuristics and biases in judgment that results is not restricted to the lay public but has been demonstrated in educated people like clinicians with training in statistical analyses. Heuristics can lead to inaccurate estimates of probability and suboptimal clinical guidance. The availability heuristic, the representativeness heuristic, and confirmatory bias are three examples that can lead to errors in clinical judgment.

Availability Heuristic

Individuals rely on past experiences to predict future events. The availability heuristic manifests when an individual estimates the probability of future events based on easily remembered past events. A doctor who diagnoses a rare disease in a patient may place a higher probability on the presence of that disease in future similar patients. For example, a family physician who discovers a lump during routine breast examination of a 35-year-old woman that is eventually revealed to be cancerous may overestimate the likelihood of breast cancer in future young women presenting with a lump on breast examination. As a result, the doctor may refer many more young women to screening mammography than if that first 35-year-old woman did not have cancer. In other words, the discovery of a cancerous lesion in a young woman is a memorable event that leads the clinician’s focus away from hard data on the prevalence of disease in women with the same characteristics. The availability heuristic leads to accurate estimation of the probability of future events only if the memorable nature of past events equates to the frequency of the events, a situation that is rarely true. This heuristic contributes to limitations in clinical decision making by leading to inaccurate assessment of probability.

Representativeness Heuristic

The representativeness heuristic manifests when a clinician allows the degree to which a patient’s symptoms resemble a disease state to signify the likelihood that the patient has the disease. Dawson and Harkes explain representativeness heuristic as instances when “resemblance is used as a quick means of assessing likelihood.” For example, the authors cite the example of a young, healthy male presenting with pleuritic chest pain, hemoptysis, and shortness of breath. The symptoms were consistent with two conditions: pulmonary infarction and bacterial pneumonitis. Representativeness bias would lead a clinician to falsely conclude that since the symptoms were consistent with two conditions that each condition was equally possible. In fact, because the symptoms don’t provide sufficient information to allow a clinician to differentiate between the two possible diagnoses, the best indicator of the correct diagnosis is the prior probability of disease. Thus, bacterial pneumonitis would have been more likely based on its prevalence in young, healthy males.

Confirmatory Bias

Individuals have the tendency to emphasize the importance of information that fits already formed hypotheses and de-emphasize information that contradicts these hypotheses. A radiologist may ignore negative
or "normal" findings on an X-ray or CT scan when they contradict his/her hypothesis that disease is present. A radiologist may seek confirmatory evidence of a hypothesis and re-interpret imaging findings in such a way as to support a preexisting hypothesis.

Values
The weight of importance that individual patients place on aspects of clinical decisions increases the complexity of clinical decision making. One patient may accept high-risk procedures such as surgery or radiation-bearing imaging examinations if there is a greater probability of gain, while another patient may avoid risk and opt to cope with disease symptoms to avoid exposure to risks of surgery or radiation. Patient values and preferences are highly variable. Clinicians may not always elicit information from patients on their values and preferences, and thus may suggest courses of action inconsistent with patient values and preferences.

How Can Decision Analysis Improve Clinical Decision Making?
Decision analysis has the potential to improve clinical decision making. Decision analysis can address and potentially overcome many of the limitations to clinical decisions identified above. By explicitly representing the features of a decision—the options, possible outcomes, associated probabilities, and values—decision analysis quantifies the relative desirability of each option based on these features. Using an explicit, quantitative framework, like that provided by decision analysis, creates the incentive to identify estimates from the best sources and specify accurate numerical risk quantities, overcoming lack of knowledge. Patient factors associated with risk, such as age, sex, and clinical findings, can be incorporated into a decision analysis with weights derived from validated risk algorithms to compute risk precisely. Indeed, validated risk algorithms can be used to calculate risk estimates based on patient characteristics. Decision analysis addresses problems of risk estimation through mathematical calculation, for example by combining the pretest probability of disease with the test result and accuracy of the test to calculate the posttest probability of disease. Inaccuracies due to heuristics and biases can be overcome by the quantitative, rule-based nature of decision analysis. Decision analysis addresses uncertainty by allowing the analyst to vary inputs within a plausible range and test changes in the desirability of each decision option. Patient values can be incorporated into a decision analysis to identify the best option.

Decision Tree
A decision tree is both a model of a decision that facilitates decision analysis, and also a tool that represents key decision features such as the decision options, possible events, probabilities, and outcomes. A decision tree is frequently represented graphically, and the decision tree features can be used to illustrate the rules associated with decision analysis. The decision tree has the following explicit components:

- A list of options to choose from
- Pathways to specify the sequence of events that can occur with each option
- Probabilities or other parameters associated with each event
- Outcomes representing the summation of a pathway
- Values indicating the desirability of the outcomes

The decision tree has the following implicit components:

- An individual (or group) whose perspective is represented by the tree. The options, pathways, and values are relevant to the individual (or group)
- A time frame that captures relevant events

Each explicit component of a decision tree is typically represented graphically. While the implicit components are not represented graphically, they must be specified and understood to facilitate interpretation of the tree.
We will illustrate the construction and analysis of a decision tree using a simple, nonclinical example. The father of an image-conscious 12-year-old girl has warned her that the forecast indicates a 40% chance of rain in the morning. He phones her from his workplace and reminds her to wear her rain slicker on her walk to school. The girl is reluctant to wear her rain slicker because she thinks it is uncool. The students in her class view rain slickers as childish. Worse yet, her rain slicker has a picture of a cartoon character on it and her father refuses to buy her another one. On the other hand, the last time she got wet on her walk to school, she felt uncomfortable in class the entire morning. The decision of whether or not to wear a rain slicker can be represented by a decision tree (Fig. 8.1).

**Constructing a Decision Tree**

**Decision Node**

The list of options to choose from is specified at the decision node, which is represented by a square. Each option is represented by a “branch” extending from the decision node. In this example, the 12-year-old’s options are “Wear the Rain Slicker” and “Leave the Rain Slicker.” The decision node is often referred to as the root of the decision tree.

**Chance Node**

In this decision, because we are uncertain about whether or not it will rain, we represent the uncertainty associated with the occurrence of these chance events. Thus, the branches for each possible chance event, “Rain” and “No Rain,” extend from a chance node, which is represented by a circle. At each chance node, the probability that the event occurs is represented along with the probability that the event does not occur. All of the events extending from a chance node are required to be mutually exclusive.

The numbers indicating the probability of each chance event are specified along the branches that extend from each chance node. Each probability is a number from 0 to 1. The probability of nonoccurrence of the event is complementary. In other words, if the probability of an event is 0.4, the probability that the event does not occur should equal 1 – 0.4 or 0.6. The sum of probabilities along the branches extending from a chance node must total 1. According to the weather forecast, the
probability of rainfall during the walk to school is 40%. Thus, the probability of no rainfall is 60%.

**Terminal Node**

The triangular nodes at the end of a pathway or a series of chance nodes are known as **terminal nodes**, because they represent the end of a sequence of events. Terminal nodes are typically represented by a triangle. Outcomes are specified at the terminal nodes. Numbers that indicate the desirability or lack of desirability of each outcome can also be represented at the terminal nodes. In clinical decision analysis, these numbers can represent life years accumulated over the time frame or a rating indicating the desirability of the health outcome. In this stylized example, we will assign a rating out of 10 for each outcome on behalf of the 12-year-old girl. If she wears the rain slicker, and there is no rain, she would feel embarrassed, and thus a ranking of 5 out of 10 seems appropriate. If she wears the rain slicker and it does rain, although her outfit will embarrass her, the relief of keeping dry on the walk to school means she might rate “Embarrassed but Dry” slightly higher at 7 out of 10. If she leaves the rain slicker and it does rain, she would feel wet and uncomfortable for most of the day, and “Wet and Uncomfortable” warrants the lowest rating of 4 out of 10. If she leaves the rain slicker and there is no rain, this would be the best outcome of all, “Cool and Dry,” earning a rating of 10 out of 10.

**Nongraphical Aspects**

It is important to note the implicit aspects of the decision that are not graphically represented by the tree. First, the perspective is that of the 12-year-old girl. The father may have different ratings for each of the outcomes. Unencumbered by feelings of embarrassment, the father would likely rate the outcome of wearing the rain slicker when it doesn’t rain much higher than 5 out of 10. Conducting the decision analysis from the father’s perspective is highly likely to change the optimal decision. Second, the time frame for the decision is not represented graphically. All of the events modeled in the tree would occur by the time the girl arrived at school. Alternatively, it is possible to build a tree to represent a longer time frame. We could model the decision to wear a rain slicker each day for the next month. The probability of rain on at least one morning for the next month would increase, potentially changing the optimal decision.

---

**Analyzing a Decision Tree**

**Expected Value**

The optimal choice in a decision analysis is one that maximizes expected value. The option with the highest expected value is the option most likely to lead to the greatest payoff, based on the events, probabilities, outcomes, and values specified. Multiplying the probability of each possible event by the value associated with each outcome allows us to calculate expected value.

**Expected Value Calculation**

To analyze the 12-year-old girl’s decision tree we would complete the following steps (Table 8.1):

1. Calculate the expected value of the option “Wear the rain slicker” by taking the average of the values for each possible outcome weighted by the probabilities.
   a. For rain, multiply the probability of rain (0.4) by the value associated with the outcome “Embarrassed but dry” (7).
   b. For no rain, multiply the probability of no rain (0.6) by the value associated with the outcome “Dry and embarrassed” (5).
   c. Add the resulting products for each possible event.
2. Similarly, calculate the expected value of the option “Leave the rain slicker.”
   a. For rain, multiply the probability of rain (0.4) by the value associated with the outcome “Wet and uncomfortable” (4).
b. For no rain, multiply the probability of no rain (0.6) by the value associated with the outcome “Dry and cool” (10).

c. Add the resulting product for each possible event.

3. Choose the option with the highest expected value.

In this case, “Wear the rain slicker” is associated with an expected value of 5.8, and “Leave the rain slicker” is associated with an expected value of 9.6. Thus, the best option, based on the 12-year-old’s values, would be to leave the rain slicker, as this option maximizes the expected rating of the quality of her day.

Calculating expected value with a decision tree is colloquially referred to as “rolling back” the tree. Since expected value calculations involve multiplication, the order of operation does not change the results. However, it is often easier to conceive of multiplication “rolling back,” proceeding from the terminal nodes to the root node. Expected value calculations are based on the concepts of randomness and long-run averages. A thought experiment of making the same choice many times underlies expected value calculations. If the 12-year-old girl made the same decision to leave the rain slicker each day for 100 days (with all other elements remaining the same), the average rating over 100 days would be 9.6.

Rationality

One criterion for optimal decision making is rationality. Expected utility theory specifies that a rational decision is one that maximizes expected value, and is consistent with the values of the decision maker. Many studies demonstrate that human beings frequently depart from “rationality” and choose options that do not maximize expected value. The reasons for this are complex. Some of the cognitive limitations discussed earlier in the chapter play a role. Additional factors such as risk aversion and the way decisions are framed also play a role. In behavioral economics, normative approaches to the study of decision making refer to the study of what individuals ought to choose based on maximizing expected value. Descriptive approaches to the study of decision making focus on what individuals actually choose and provide explanations for departures from rational choices. Well-established psychological characteristics of humans, such as aversion to loss and sensitivity to how options are framed, are discussed further in work by Tversky and Kahneman, among others.15

Limitations

Even this stylized example of a decision tree has important limitations that also threaten the thoroughness and validity of more complex decision analyses. The tree may exclude reasonable options. The 12-year-old could carry the rain slicker and put it on only if it began to rain. We could then incorporate ratings on the desirability of getting a little wet in the moments before putting on the rain slicker. She might also consider bringing an umbrella. The ratings indicating the desirability of each outcome were not elicited directly from the 12-year-old girl. Based on an understanding of the girl’s preferences, we as decision analysts assigned ratings. The 12-year-old’s ratings may indeed be different. It is important to note that the process of assigning ratings is analogous to what occurs in a clinical encounter. Clinicians make recommendations to patients on the basis of clinical effectiveness and informed by what they understand to be the patient’s preferences and values.

We have not explored the effect of uncertainty on the optimal decision. In this

---

### Table 8.1 Expected value calculations for rain slicker decision options

<table>
<thead>
<tr>
<th>Possible events</th>
<th>Probability × value</th>
<th>Expected value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wear the rain slicker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain</td>
<td>0.4 × 7 = 2.8</td>
<td>2.8 + 3.0 = 5.8</td>
</tr>
<tr>
<td>No rain</td>
<td>0.6 × 5 = 3.0</td>
<td></td>
</tr>
<tr>
<td>Leave the rain slicker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain</td>
<td>0.4 × 4 = 3.6</td>
<td>3.6 + 6.0 = 9.6</td>
</tr>
<tr>
<td>No rain</td>
<td>0.6 × 10 = 6.0</td>
<td></td>
</tr>
</tbody>
</table>
example, if the probability of rain were 70%, the optimal decision would change to “Wear the rain slicker.” Uncertainty analysis is explored in more depth in the next chapter.

■ **Key Aspects of Clinical Decision Analysis**

**Perspective**

The perspective signals what options, events, and outcomes are relevant to incorporate in the decision analysis. Clinical decision analysis can be conducted from the perspective of an individual patient or a group (often called a cohort) of patients. A cohort should be a homogeneous group with characteristics defined by age, sex, condition severity, disease stage, or setting.

**Options**

The options in a clinical decision analysis correspond to options important in clinical practice. Therapeutic interventions, screening modalities, diagnostic interventions, medications, surgeries, or some combination of these may be of interest. Programs such as quality improvement or screening schedules with varying frequency may also be included as options. The option to do nothing is a valuable component of a clinical decision analysis. For many clinical questions a “watch and wait” strategy is optimal. Furthermore, including a do nothing option often forces the analyst to think of the natural history of disease, in such a way that forms the basis for other strategies.

**Parameters and Probabilities**

A range of probabilities and other parameters are required for clinical decision analysis. Parameters for the test characteristics, such as the sensitivity and specificity, play an important role when considering diagnostic interventions. Decision analysis involving diagnostic tests can represent the interaction of pretest probability and diagnostic accuracy by explicitly incorporating prevalence data for the population of interest to denote pretest probability of disease. The probabilities associated with therapeutic success and the probabilities of adverse effects are also important in clinical decision analysis.

**Outcomes**

Outcomes such as survival are commonly incorporated into clinical decision analysis either as a measure of life expectancy or as a measure of the proportion of patients surviving for a period of time. The quality-adjusted life-year (QALY) is an outcome measure that expresses the quantity of life weighted by the quality of life. The QALY will be explored in depth in the next chapter on economic evaluation. Outcomes are also commonly expressed in natural units, such as diagnostic accuracy, the number of cancer cases detected, or the number of patients correctly diagnosed.

**Time Frame**

The time frame for a clinical decision analysis should be long enough to capture important events. If the decision analysis pertains to an acute illness that resolves quickly and has no long-term effects, then a short time frame, for example, 1 year may suffice. However, this is rarely the case with clinical decisions. Interventions typically improve long-term prognosis or introduce the probability of long-term harms. As a result, longer time frames are often preferred.

**Values**

Values are not always incorporated into clinical decision analysis. The implicit values inherent in the outcome provide insight into interpreting clinical decision analyses. For example, correctly diagnosing more patients is generally preferred to correctly diagnosing fewer patients and increasing survival is preferred over decreased survival. One individual may value an active lifestyle more than another individual and this could lead to differences in clinical decision making. As is demonstrated in the next chapter, it is possible to assign values to outcomes using utilities to calculate QALYs.

**Implementation of Decision Trees**

As demonstrated with our simple example, a decision tree can be implemented using pencil
and paper. Clinical decision analyses are typically more complex, and a variety of software programs can help with managing the complexity. Tree Age Professional provides a graphical, user-friendly interface with a range of analytic tools. Performing sensitivity analysis is the best way to assure the internal validity of the decision tree. All inputs should be varied one at a time over extreme and possibly implausible ranges in a process known as “debugging.” For example, all probabilities should be varied from 0 to 1. If decision tree results are not consistent with logical predictions, there is a strong indication of a “bug” or error in the programming. For example, it would be suspicious if a decision tree indicates a therapy is the best option even after setting the probability associated with the effectiveness of that therapy to 0. Sensitivity analysis is also important to understand how the optimal decision would change if tree inputs such as probabilities were varied over a range that represents uncertainty about the exact value. In one-way sensitivity analysis, each tree input is varied one at a time, while all other inputs are held constant. Sensitivity analysis is explored in greater depth in the next chapter.

**Sensitivity Analysis**

Sensitivity analysis, also called uncertainty analysis, is the process of systematically varying decision tree inputs to assess changes in the results. Performing sensitivity analysis is the best way to assure the internal validity of the decision tree. All inputs should be varied one at a time over extreme and possibly implausible ranges in a process known as “debugging.” For example, all probabilities should be varied from 0 to 1. If decision tree results are not consistent with logical predictions, there is a strong indication of a “bug” or error in the programming. For example, it would be suspicious if a decision tree indicates a therapy is the best option even after setting the probability associated with the effectiveness of that therapy to 0. Sensitivity analysis is also important to understand how the optimal decision would change if tree inputs such as probabilities were varied over a range that represents uncertainty about the exact value. In one-way sensitivity analysis, each tree input is varied one at a time, while all other inputs are held constant. Sensitivity analysis is explored in greater depth in the next chapter.

**Simple Decision Tree Using a Diagnostic Testing Strategy**

A decision tree incorporating a choice of diagnostic testing strategies follows a particular structure. An example is provided in Fig. 8.2,

---

**Fig. 8.2** CT angiography is weighed against conventional angiography for diagnosis of intracranial aneurysm in subarachnoid hemorrhage. Explanation is provided in the accompanying text. Abbreviations: CT, computed tomography; IV, intravenous; IA, intraarterial.
where different imaging strategies (CT angiography versus conventional catheter angiography) are compared for the diagnosis of intracranial aneurysm in a patient presenting with subarachnoid hemorrhage. The list of diagnostic testing options under consideration is specified at the decision node. At the first chance node, we incorporate the probability that the disease for which we are testing is present—in other words, the pretest probability of disease. In this example, we are testing for the presence of an intracranial aneurysm, and so the first chance node reflects the probability of having an underlying intracranial aneurysm in a patient with subarachnoid hemorrhage, which has been reported as approximately 85%. At the next chance node, we incorporate the test results, with the option of the test result being either positive or negative.

Based on whether the underlying disease is present or not, and whether the test results are positive or negative, we have four possible scenarios that may ensue for the results of each testing strategy: true positive, false negative, false positive, or true negative. Considered this way, it becomes very straightforward to incorporate known test characteristics derived from the literature into the decision tree. The probability of having a true positive result is equivalent to the sensitivity of the test, the probability of having a false negative result is equivalent to \((1 – \text{sensitivity})\) of the test, the probability of having a false positive result is equivalent to \((1 – \text{specificity})\) of the test, and the probability of having a true negative result is equivalent to the specificity of the test. Terminal nodes and outcome values are specified in a similar manner to other decision trees. Outcome values may take into account the desirability of having a certain test (e.g., a non-invasive test is more desirable than an invasive test) or a certain result (e.g., a true diagnosis is more desirable than a missed or false diagnosis).

Input values are specified in Table 8.2. In the example provided, CT angiography is preferred to conventional angiography with a higher overall expected value (0.88 vs. 0.79) (Table 8.3).

### Example from the Published Literature

In this clinical decision analysis, the authors considered MRI and PET/CT for imaging patients with stage IB clinically inoperable cervical cancer in order to improve cancer staging prior to initial therapy. Detecting advanced cancer stage earlier as indicated by parametrical extensions of the tumor or lymph node involvement

---

**Table 8.2 Model inputs**

<table>
<thead>
<tr>
<th>Model inputs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of aneurysm in subarachnoid hemorrhage</td>
<td>85%(^1)</td>
</tr>
<tr>
<td>Test information</td>
<td></td>
</tr>
<tr>
<td>Sensitivity of CTA</td>
<td>95%(^2)</td>
</tr>
<tr>
<td>Specificity of CTA</td>
<td>96%(^2)</td>
</tr>
<tr>
<td>Sensitivity of conventional angiography</td>
<td>99%(^3)</td>
</tr>
<tr>
<td>Sensitivity of conventional angiography</td>
<td>99%(^3)</td>
</tr>
</tbody>
</table>

Outcome values:

1. True diagnosis but IV contrast and radiation
2. Missed diagnosis and IV contrast and radiation
3. False diagnosis and IV contrast and radiation
4. True diagnosis but IV contrast and radiation
5. True diagnosis but invasive examination, IA contrast, and radiation
6. Missed diagnosis and invasive examination, IA contrast, and radiation
7. False diagnosis and invasive examination, IA contrast, and radiation
8. True diagnosis but invasive examination, IA contrast, and radiation

\(^*\) Probability of disease and test information are derived from the literature; outcome values are derived from the literature (if available) or estimated from expert opinion. Abbreviations: CTA: computed tomography angiography; IV, intravenous; IA, intraarterial.
can lead to better targeting of therapies. In particular, the most aggressive therapy trimodality therapy (surgery, chemotherapy, and radiation) is associated with high rates of morbidity. Thus, better targeting can avoid problems of overtreatment and undertreatment and improve patient outcomes.

**Perspective**

The decision analysis is conducted from the perspective of a hypothetical cohort of patients with FIGO (International Federation of Gynecology and Obstetrics) IB clinically operable cervical cancer. In contrast to the previous examples, in which we performed an analysis of decision making for one individual, the goal of this decision analysis was to inform decision making about a cohort, a group of clinically similar individuals.

**Options**

The authors considered four possible strategies: (1) MRI alone; (2) PET/CT alone; (3) combined MRI and PET/CT; and (4) no further staging.

**Parameters and Probabilities**

The authors incorporated the following probabilities into the decision analysis, where necessary, converting rates into the probabilities required by a decision tree.

- Pretest probabilities of advanced cancer stage (parametrial extensions, lymph node involvement)
- MRI sensitivity
- MRI specificity
- PET/CT sensitivity
- PET/CT specificity
- Operative mortality rates for modified radical hysterectomy
- Radiation fatality rates

**Outcomes**

The three outcomes of interest in the clinical decision analysis were: (1) the percentage of patients surviving 5 years following diagnosis; (2) the percentage of patients receiving the correct primary treatment based on the
true extent of disease; and (3) the percentage of patients spared trimodality therapy.

**Time Frame and Values**

The time frame for the decision analysis was 5 years; no values were assigned to the outcomes beyond the implicit values associated with the outcomes of survival, correct diagnosis, and avoidance of overtreatment.

**Implementation of Decision Tree**

The decision tree was implemented in the software TreeAge Professional 2007 (TreeAge Software).

**Summary of Results**

The authors found that all strategies were associated with similar rates of 5-year survival. PET/CT resulted in the highest proportion of patients receiving correct primary treatment. The combined MRI with PET/CT strategy resulted in the greatest percentage of patients avoiding trimodality therapy. In sensitivity analyses, the authors indicate that varying the sensitivity and specificity of the PET/CT and MRI scans resulted in changes to the optimal strategy. For example, with very high specificity and low sensitivity for PET/CT, MRI was preferred for avoiding trimodality therapy.

**Related Topics**

**Clinical Decision Aids**

Clinical decision aids are tools that exist in a variety of formats (pamphlets, videos, and websites) that help patients clarify their values when making clinical decisions in addition to providing information about options and outcomes. Clinical decision aids are informed by similar theoretical frameworks to decision analysis including the economic concepts of values and expectations, general psychology, in addition to decisional conflict, social support, and social psychology frameworks. Systematic review of the literature demonstrates that the use of clinical decision aids improves the quality of clinical decision making including reduced decisional conflict and decreased passivity in decision making.

**Clinical Prediction Rules**

Clinical prediction rules are structured tools that allow clinicians to quantify risk by assigning weights to different risk factors. Incorporation of risk algorithms into clinical practice can lead to improved clinical decision making. A randomized trial in which decision support systems provided quantitative pretest probabilities to clinicians reduced unnecessary medical radiation exposure in emergency department patients with chest pain and dyspnea. Use of clinical prediction tools has been facilitated by technology. For example, they have been incorporated into electronic medical records medication order systems, or made available online. Adjuvant online is a clinical prediction tool for the risk of breast cancer recurrence in women diagnosed with early stage breast cancer. However, risk calculators require updating in response to newly identified factors. Clinicians may neglect to assess important emerging risk factors and modify risk estimates accordingly. Clinical prediction tools have varying quality and some are not well validated.

**Cost-Effectiveness Analysis**

Cost-effectiveness analysis is the comparative assessment of options for health care intervention in terms of both costs and outcomes. The key metric in a cost-effectiveness analysis is the incremental cost-effectiveness ratio, which signifies the incremental cost for an incremental improvement in a health outcome. Cost-effectiveness analyses frequently employ decision analytic techniques to inform health policy decision making. Clinical decision analysis usually excludes consideration of costs. While clinical decision analysis has had limited uptake in real-world clinical practice, cost-effectiveness analysis is increasingly used in health policy decision making to identify optimal avenues for investing in health care. Cost-effectiveness analysis is discussed...
in more detail in the next chapter, along with other types of economic evaluation.

■ Highlights of Key Points

• Clinical decisions are challenging for patients and clinicians alike due to the complexity and volume of available information.
• Heuristics are shortcuts that produce errors in judgment and lead to suboptimal decisions.
• Decision analysis represents probabilities, outcomes, and values in a tree structure and combines them mathematically in order to identify the option with the largest expected value.
• The principles of decision analysis form the basis for decision aids, clinical decision support tools, and cost-effectiveness analysis.

References


Costs and Consequences: Economic Evaluation in Radiology

Wendy J. Ungar and Richard M. Zur

Learning Objectives

- To provide a basic understanding of the principles of economic evaluation.
- To describe the elements in a cost-effectiveness analysis (CEA).
- To facilitate understanding of economic evaluation scientific literature.
- To illustrate CEA with examples in the field of radiology.

Introduction

Why Economic Assessment in Health Care?

Should radiographs be performed routinely for patients presenting in the emergency department with ankle sprains? Should health insurance plans raise premiums to pay for more frequent cancer screening? Should a publicly financed government health plan pay for an expensive experimental treatment that may or may not extend the life of a terminal patient with a rare genetic disorder? When does it make sense for hospitals to invest in more beds and nurses?

These are the kind of questions that health care systems grapple with on a daily basis. There is not an unlimited supply of money to pay for all the health care needs and wants of a population. Health resources such as the number of specialists, number of hospital beds, or number of magnetic resonance imaging (MRI) scanners are limited by the budget that pays for them. All of the programs, services, and treatments that result in health benefits for members of society are competing for the same scarce dollars. An efficient allocation of finite health care dollars therefore requires that a choice be made: Will it be more nurses or more radiographs?

It is the goal of health economic evaluation to determine which option is the one that is most economically efficient—that is, that maximizes health gains per dollar spent and achieves value for money. Every dollar spent on a particular program or service takes a dollar away from the next best possible use of that money. Poor allocation of scarce health care dollars results in waste and is represented by the opportunity cost associated with the choice made.

In this chapter the basic concepts and methods of economic evaluation are introduced and illustrated as they pertain to the field of radiology.

Why Economic Assessment in Radiology?

Perhaps more than most medical specialties, the field of radiology is faced with economic pressures. The acquisition of imaging devices, the use of those devices for screening or diagnosis, the use of radiotherapy in oncology, the personnel dedicated to operation, maintenance of radiotherapy and imaging equipment, and the need to upgrade and expand use are common considerations in the life of a radiologist.¹

Once reserved for specialized consultation in rare cases, MRI has become routine in the diagnosis of a wide range of disorders. There were an estimated almost 2 million MRI exams performed in Canada in 2014 to 2015.² This is 20 times more than the number performed in 1990. One might wonder if this increased use has resulted in a concomitant reduction in the use of alternative imaging approaches such as computed tomography (CT).
In fact, 5.3 million CT exams were performed on Canadian patients in 2014–2015, 1.8 times more than the number of exams performed in 1990. In the same year, the rate of MRI exams performed ranged from a low of 33 per 1,000 persons in British Columbia to a high of 55 per 1,000 persons in New Brunswick. CT exams ranged from 91 per 1,000 persons in Alberta to 174 per 1,000 persons in New Brunswick.

These statistics prompt a number of questions for the budding radiologist-health economist. Does the relatively high rate of imaging in New Brunswick result in better health outcomes for residents of this province compared to residents of other provinces? What is the opportunity cost of imaging in New Brunswick? Is less money being spent on other health services resulting in poorer health outcomes in other clinical areas? Are radiologists in Canada performing an MRI when a less expensive CT might suffice? Are they ordering excessive imaging exams? Are choices between different imaging approaches made with evidence of value for money?

All of these questions are relevant to the field of economic evaluation. Indeed, questions of economic efficiency apply not only to radiation therapy and diagnostic radiology, but to population screening such as mammography to identify breast tumors or bone density screening to identify women at risk for osteoporosis. With an aging population, the demand for radiation therapy for cancer treatment and imaging for screening and diagnosis is expected to increase, underscoring the critical importance of making the wisest allocation decisions possible to ensure the maximum health benefits are achieved for the money spent.

In summary, the goals of economic evaluation are:

- To increase efficiency in the selection of treatments/diagnostic strategies
- To promote optimal use of available therapies
- To understand the relationship between health care system policies that invest in health interventions, services and programs, and patient health outcomes
- To enhance the health status of target populations
- To achieve a net welfare gain in society through improved health of the population

It is equally important to understand that economic evaluation is not a form of cost containment or cost reduction nor is it a specific health policy. Health economic evaluation involves the use of rigorous methods to produce evidence to facilitate decisions regarding investment in health care resources. The ultimate goal of health economic evaluation is to inform health care decision making so that the greatest health benefits can be achieved for any given investment.

### Concepts

#### Defining Economic Evaluation

An economic evaluation is defined as a comparison of two or more interventions in terms of both costs and health consequences (Fig. 9.1).

In the comparison of interventions, the inputs are the costs associated with health care resources consumed. Health care resources may include physician consultations, imaging tests, technician time, surgical procedures, and use of medical devices. These are weighed against the outputs, which are the changes in health status that are a consequence of using the health interventions of interest. Measuring the volume of resources, such as the number of imaging tests performed, is never the endpoint in an economic evaluation. Health resources are consumption items that must always be weighed against the output or health consequences, which are the observed changes in health status.

The interventions being evaluated are often a new or emerging treatment that is compared to standard care, or they may be an array of possible approaches compared to standard or usual care. Sometimes an intervention represents an entirely novel treatment for a condition for which no treatment previously...
cost–minimization analysis (CMA). When comparing two or more interventions, one must collect all of the resources used (costs) and all of the health status changes (consequences) for all the interventions being compared, regardless of the analytic approach. All four approaches measure health care resource use in terms of monetary units, such as dollars. The four approaches differ in terms of how health outcomes are specified.

There are four different types of analytic approaches to economic evaluation. These four approaches are listed and defined in Table 9.1.

The four types of economic evaluations are cost-effectiveness analysis (CEA), cost–utility analysis (CUA), cost–benefit analysis (CBA), and cost–minimization analysis (CMA). When comparing two or more interventions, one must collect all of the resources used (costs) and all of the health status changes (consequences) for all the interventions being compared, regardless of the analytic approach. All four approaches measure health care resource use in terms of monetary units, such as dollars. The four approaches differ in terms of how health outcomes are specified.

The most common type of economic evaluation, CEA, measures health outcomes in natural health units. Examples include how many people live or die, the number of years of additional life achieved for each intervention, or the number of adverse events or complications observed for each intervention.

Table 9.1  Types of economic evaluations

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Measurement of costs</th>
<th>Measurement of consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost-effectiveness</td>
<td>Resource use expressed in monetary units</td>
<td>Single effect, common to alternatives, e.g., life years</td>
</tr>
<tr>
<td>Cost-utility</td>
<td>Resource use expressed in monetary units</td>
<td>Multiple attributes combined into a common universal metric, e.g., quality-adjusted life years (QALYs)</td>
</tr>
<tr>
<td>Cost–benefit</td>
<td>Resource use expressed in monetary units</td>
<td>Health benefits converted to monetary units</td>
</tr>
<tr>
<td>Cost–minimization</td>
<td>Resource use expressed in monetary units</td>
<td>None—evidence reveals no clinically important differences between alternatives</td>
</tr>
</tbody>
</table>
The precise choice of health status measure will usually vary between medical specialties and often varies from study to study. This can complicate comparisons between studies within a therapeutic area and will make it difficult, if not impossible, to make allocation decisions across therapeutic areas. Should one spend more money on mammography (number of cases of breast cancer detected) or treatment of depression (number of patients achieving good mental health)? CEA is explained in greater detail in the cost-effectiveness analysis section of this chapter.

The problem of disparate outcomes is solved by using a common metric. In CUA the quality-adjusted life year (QALY) is used. The QALY is a composite measure that considers not only the life years achieved for any given intervention, but also the health-related quality of life that the patient experiences during those life years. Thus, while a CEA might indicate that a liver transplant and medical treatment of advanced breast cancer may result in the same number of life years for the two patient groups, only a CUA can reveal that the surgical patients also have a much better quality of life. The preference for the higher quality of life state, or utility for the health state, is used as weight to adjust the observed life expectancy. Knowing that there is an improvement in quality-adjusted life years, not just the number of life years, is an important consideration when making an allocation decision. A significant advantage of the QALY is that it allows comparisons not only across different studies, but across very different patient groups and therapeutic areas. For that reason it is considered a universal outcome measure and is recommended by economic evaluation guidelines. Despite the allure of the universal metric, CUAs are often difficult to carry out as available health-related quality of life measures may not have the appropriate performance characteristics or sensitivity to detect changes in health states in some patient populations.

The terms “cost–benefit” or “cost–benefit analysis” are often used by lay persons to vaguely denote any type of economic evaluation. This is an unfortunate misuse of the term, since CBA has a precise meaning in economic evaluation. Only a CBA attempts to convert health status changes into dollar terms. More than that, a well-executed CBA will consider nonhealth benefits, such as better workforce participation resulting from improved health, as well as benefits to other individuals (spillover effects), such as improvements in the quality of life of the patient’s family members. Because the benefits measured are so wide ranging, of the four analytic approaches, only a CBA can claim to be founded in welfare economics. That being said, the methods required to convert health effects into dollars are complex and often difficult to undertake. For that reason, there are few CBAs in health care.

The final approach listed is CMA. In a cost-minimization analysis only the costs are compared between two or more interventions. This is not because the outcomes are not considered relevant or important to the research question, but because existing high quality, reliable evidence demonstrates that the two interventions may result in comparable levels of effectiveness. One cannot proceed with a CMA based on the assumption of equivalent health status improvement—there must in fact be evidence of the case. In reality, few studies are labeled as CMAs because although the observed difference in health status between two interventions may be below the threshold deemed clinically important, the difference may nevertheless have relevance for the weighing of costs against health benefits. CMA remains the least common of the analytic approaches.

The fundamental defining concept that an economic evaluation compares two or more interventions in terms of costs and health consequences is often illustrated with the use of the cost-effectiveness plane. As seen in Fig. 9.2, costs are plotted on the y-axis and health effects on the x-axis. The plane is divided into four quadrants, which can be labeled for convenience as the compass points northeast, southeast, southwest, and northwest.

If compared to standard care a new intervention is more costly and not as effective (lies in the northwest quadrant), then it is dominated by standard care and is clearly not a worthy investment. If the intervention
157

Costs and Consequences: Economic Evaluation in Radiology

The Cost-Effectiveness Plane

Compared to standard care, the intervention is:

- **Northwest**: Standard is DOMINANT (Lose-Lose)
- **Northeast**: More Costly/More Effective
- **Southwest**: Less Effective/Less Costly
- **Southeast**: Intervention is DOMINANT (Win-Win)

Fig. 9.2 The cost-effectiveness plane. Compared to standard care, a new intervention can be more costly but more effective (northeast quadrant), less costly and more effective (southeast quadrant), less costly and less effective (southwest quadrant), or more costly and less effective (northwest quadrant).

Costing

An economic evaluation weighs the additional costs of a new intervention against any added savings money and results in better health outcomes compared to standard care (lies in the southeast quadrant), then it is surely worth adopting. The remaining two quadrants are the ones for which economic evaluation is most critical: when the intervention is less costly but also less effective (southwest quadrant) or when the intervention is more costly but also is more effective than standard care (northeast quadrant). In a majority of cases, detailed economic evaluations are carried out because the novel intervention is more costly but promises better effectiveness compared to standard care. The salient questions are: How much more costly? How much more effective? Is the added cost worth the added health benefit? This is the crux of economic evaluation. The basic methods are further delineated in the sections to come.

The Research Question

The research question of an economic evaluation should flow from a problem identified by a relevant stakeholder such as clinical or policy decision-maker and should address a gap in existing evidence. Economic evaluations are often motivated by the introduction of a new technology that is more expensive but also promises to be more effective than current approaches to care. Enhancements to CT scanning and MRI often meet these criteria. In formulating the research question, the researcher should also consider the target audience for the study. Begin by:

- Identifying a gap in the evidence base
- Stating the study purpose and rationale
- Identifying the target knowledge user audience
- Posing a clear research question in terms of costs, outcomes, comparators, perspective, and target population

The example in Box 9.1 illustrates how to define a research question in an economic evaluation with an example from diagnostic radiology.6
Research Methods in Radiology

Box 9.1. Constructing an economic evaluation research question

Acute appendicitis is common and failure to diagnose it in a timely and accurate fashion can expose a child to the risk of serious infection with high morbidity and costly hospital stays. An accurate and timely diagnosis can be facilitated with the use of CT in addition to or instead of ultrasound. However, CT may expose the child to unnecessary and potentially harmful radiation. CT can also be costly to operate and may not be readily available when required.

A recent study evaluated the cost-effectiveness of different imaging approaches for the diagnosis of suspected appendicitis in children. The study's purpose and rationale were first presented as follows:

- CT offers greater diagnostic accuracy compared to ultrasound, but is more costly to operate.
- CT is associated with a latent risk of malignancy and children are especially susceptible as they have more time to develop a malignancy.
- In view of this risk, pediatric hospitals have been seeking ways to reduce exposure to radiation through adjusting settings, lowering the dose, and reducing the use of CT.
- There is little published evidence regarding the effectiveness or cost-effectiveness of alternative radiological approaches to diagnosing acute appendicitis.
- Institutional decision-makers and clinical knowledge users require evidence to determine whether CT alone or CT used only when ultrasound results are inconclusive offers an advantage over ultrasound alone to inform decisions about treatment, budget allocation, and staffing.

The primary research objective was posed as follows:

The study objective was to estimate the incremental costs per quality-adjusted life year gained of CT alone or CT following an inconclusive ultrasound compared to ultrasound alone, from a third-party payer perspective in children presenting with suspected appendicitis.

benefits compared to standard care. While “benefits” may be represented in terms of a single effectiveness variable such as life years or quality-adjusted life years, the cost consists of many cost item variables, all of which must be properly combined to accurately represent the costs of an intervention. The costing process is comprised of several steps and requires extensive data collection. These steps are:

1. Cost item identification
2. Cost item measurement
3. Valuation

Cost Item Identification

The cost items that are included in an economic evaluation can be broadly grouped into three categories. These include:

1. Direct health care–related costs
2. Direct patient (out-of-pocket) costs
3. Indirect costs

One of the first steps is to consider how health care is delivered given the clinical framework set by the primary research question. Which types of health care resources are used in the delivery of care to the patient population for whom the intervention will be targeted? One might start by making a detailed list of all the potential health care resources that might be consumed.

Although not comprehensive, Table 9.2 shows examples of typical direct health care cost items that could be considered in an economic evaluation of a radiological intervention.

Few health care systems cover 100% of all the health care costs for patients. Depending on how the health plan benefits are structured, patients may incur some out-of-pocket costs to receive care regardless of whether the health care plan is publicly or privately financed. Many health plans require patients to make copayments for physician services, emergency services, ambulance services, in-patient stays, and medications. Copayments may be fixed or a percent of the total cost, depending on the cost item and the health plan benefits. In addition, patients routinely pay out-of-pocket for items that are not covered...
Out-of-pocket costs can be considerable and pose a financial burden for some patients and their families over a short or long interval.\(^8\) A large out-of-pocket expenditure burden may force patients to make choices that affect the health services used by other family members. It is thus important that these cost items not be overlooked.

The third category of cost items are the indirect costs, also referred to as productivity costs. In addition to paying out-of-pocket for some or all health care, patients, and often members of their family, miss time from work and other activities due to their illness or the need to receive care. These time losses too can be substantial. Perhaps the greatest cost stemming from time loss is due to premature death. Long-term disability due to illness also results in long-term productivity losses. Failing to account for these significant costs can lead to misclassifying an intervention that is cost saving as one that is cost-incurring. Table 9.4 lists some of the types of time losses that result in productivity costs.

It is important to note that often the patient’s family members and informal caregivers will incur indirect costs for the care of the

---

### Table 9.2 Examples of direct health care cost items

<table>
<thead>
<tr>
<th>Category</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Imaging equipment</td>
</tr>
<tr>
<td></td>
<td>Equipment depreciation</td>
</tr>
<tr>
<td></td>
<td>Overhead and maintenance contracts</td>
</tr>
<tr>
<td></td>
<td>Contrast media (i.e., gadolinium injections for contrast-enhanced MRI)</td>
</tr>
<tr>
<td></td>
<td>Digital image storage and retrieval systems (Picture Archiving and Communication System - PACS)</td>
</tr>
<tr>
<td></td>
<td>Imaging tests</td>
</tr>
<tr>
<td></td>
<td>Radiologist and trainee (fellow, resident) services</td>
</tr>
<tr>
<td></td>
<td>Radiology technician services</td>
</tr>
<tr>
<td>Inpatient care</td>
<td>Ward hotel costs</td>
</tr>
<tr>
<td></td>
<td>Hospital overhead costs (laundry, housekeeping, utilities, capital)</td>
</tr>
<tr>
<td></td>
<td>Nursing services</td>
</tr>
<tr>
<td></td>
<td>Specialist physician assessments and consultations (may vary by specialty)</td>
</tr>
<tr>
<td></td>
<td>Surgical procedures</td>
</tr>
<tr>
<td></td>
<td>Laboratory tests</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
</tr>
<tr>
<td></td>
<td>Ambulance services</td>
</tr>
<tr>
<td></td>
<td>Emergency department services</td>
</tr>
<tr>
<td>Outpatient care</td>
<td>Physician assessments (may vary by specialty)</td>
</tr>
<tr>
<td></td>
<td>Laboratory tests</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
</tr>
<tr>
<td></td>
<td>Medical devices</td>
</tr>
<tr>
<td></td>
<td>Home care services</td>
</tr>
<tr>
<td></td>
<td>Complementary health practitioner services (may vary by specialty)</td>
</tr>
</tbody>
</table>

### Table 9.3 Examples of direct patient cost items

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copayment for inpatient stays</td>
</tr>
<tr>
<td>Copayment for outpatient physician assessments</td>
</tr>
<tr>
<td>Complementary health practitioner services (may vary by specialty)</td>
</tr>
<tr>
<td>Copayment for laboratory tests</td>
</tr>
<tr>
<td>Prescription medications or copayment for medications</td>
</tr>
<tr>
<td>Copayment for ambulance services</td>
</tr>
<tr>
<td>Copayment for emergency department services</td>
</tr>
<tr>
<td>Over-the-counter medications</td>
</tr>
<tr>
<td>Complementary medications</td>
</tr>
<tr>
<td>Medical devices and appliances</td>
</tr>
<tr>
<td>Home care services</td>
</tr>
<tr>
<td>Imaging preparation materials</td>
</tr>
<tr>
<td>Rehabilitation programs</td>
</tr>
<tr>
<td>Health education programs</td>
</tr>
<tr>
<td>Travel, accommodation, and transportation to site of health service delivery</td>
</tr>
</tbody>
</table>
exactly what items are included in the bundle. Oftentimes missing items must be added in, such as inpatient physician consultation fees.

After identifying all the cost items that may be relevant to an economic evaluation, the next step is to identify the data sources that will supply information about utilization volume and price. Utilization volume for each item is usually determined at the patient level. Some prices may also be determined at the patient level, but typically prices for health care goods and services are fixed, especially in publicly financed health care systems. Fixed prices may vary over time or across jurisdictions.

It is not unusual to use multiple data sources for any given study. It is also often useful to consider multiple data sources for each of the variables related to volume and price—as long as the data are accurate, reliable, and are appropriate for the research question.

Cost item data for economic evaluations can be collected retrospectively, prospectively, or both. Researchers usually begin by determining what patient-level resource use data can be obtained from existing sources. Common sources for retrospective data collection of cost item utilization volume include patient charts and health plan administrative databases. The latter can serve as valuable sources for stable estimates of health resource use such as frequency of admissions, number of doctor visits, and number of medication prescriptions within a given study period. Other sources for retrospective data collection include patient registries and existing clinical study databases.

When resource use data are not readily available from existing sources, or if the data quality is low, a prospective study may be undertaken for some or all necessary cost items. An example of a cost item that is rarely documented and might need to be collected prospectively is time losses associated with patient care. Prospective data collection may include simple surveys or questionnaires. When the bulk of cost (and outcomes) data are to be collected prospectively for the experimental intervention and the standard care comparators, then epidemiologic principles of study design should be followed (Chapters 5 and 6). Ideal study designs include randomized controlled

### Cost Item Measurement

For every cost item listed in the tables, there are actually two variables that are relevant: the utilization quantity and the price. It becomes clear that data collection for costing is a significant task. Cost item measurement consists of collecting the necessary data related to both utilization quantity and price for each item. Some cost items themselves may be comprised of multiple items. As seen in Table 9.2, the cost of a hospital admission has been broken down into its component items. Institutions that employ case costing management information systems may be able to supply a bundled hospital admission cost for the type of patients being studied. In these situations it is important to ascertain

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence from paid labor</td>
<td>Days lost from paid employment due to illness.</td>
</tr>
<tr>
<td>Absence from unpaid labor, volunteer work, or usual activities</td>
<td>Days lost from unpaid work or volunteer work due to illness.</td>
</tr>
<tr>
<td>Restricted activity days</td>
<td>Days lost due to illness or injury.</td>
</tr>
<tr>
<td>Lost leisure time</td>
<td>Days lost due to exhaustion or leisure activities.</td>
</tr>
<tr>
<td>Informal caregiver time</td>
<td>Days lost due to caring for ill family members.</td>
</tr>
<tr>
<td>Travel time to access care</td>
<td>Days lost due to transportation to medical care.</td>
</tr>
<tr>
<td>Waiting time</td>
<td>Days lost due to waiting for medical care.</td>
</tr>
<tr>
<td>Long-term disability</td>
<td>Years of potential life lost due to illness.</td>
</tr>
<tr>
<td>Premature death</td>
<td>Days of life lost due to illness.</td>
</tr>
</tbody>
</table>

### Table 9.4 Examples of indirect cost items

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence from paid labor</td>
<td>Days lost from paid employment due to illness.</td>
</tr>
<tr>
<td>Absence from unpaid labor, volunteer work, or usual activities</td>
<td>Days lost from unpaid work or volunteer work due to illness.</td>
</tr>
<tr>
<td>Restricted activity days</td>
<td>Days lost due to illness or injury.</td>
</tr>
<tr>
<td>Lost leisure time</td>
<td>Days lost due to exhaustion or leisure activities.</td>
</tr>
<tr>
<td>Informal caregiver time</td>
<td>Days lost due to caring for ill family members.</td>
</tr>
<tr>
<td>Travel time to access care</td>
<td>Days lost due to transportation to medical care.</td>
</tr>
<tr>
<td>Waiting time</td>
<td>Days lost due to waiting for medical care.</td>
</tr>
<tr>
<td>Long-term disability</td>
<td>Years of potential life lost due to illness.</td>
</tr>
<tr>
<td>Premature death</td>
<td>Days of life lost due to illness.</td>
</tr>
</tbody>
</table>

patient. This is particularly true for pediatric and other dependent patient populations. The indirect costs of family members and other informal caregivers should be measured and can be ascribed to the patient for the purpose of the analysis.

In addition to the above cost categories, there are also intangible costs.5 These are the psychosocial costs, such as decrements in quality of life or well-being, that are difficult to measure but nevertheless play an important role in understanding the burden on the patient. In the field of radiology, intangible psychosocial costs may consist of heightened anxiety and concern stemming from waiting for test results and from receiving positive test results.
trials (RCTs). Nonexperimental observational study designs, such as cohort studies, are sometimes better suited, such as when the intervention is deployed at the institutional rather than the patient level. Such would be the case in an economic evaluation comparing, for example, the use of positron emission tomography (PET)-CT for diagnosis of Hodgkin lymphoma at one hospital or in one region compared to standard CT at another. It is also expedient to consider “piggy-backing” an economic evaluation onto a planned RCT. Adding a health resource use questionnaire onto a single or multicenter RCT that is already collecting outcomes data can be a highly efficient means of data collection. However, if the RCT has limited external validity because of patient inclusion and exclusion criteria, then the health economic data that are collected may not be meaningful for decision-makers who must consider cost-effectiveness for heterogeneous patient populations. In addition, a rigid RCT protocol may introduce protocol-driven health resource costs, or may interfere with the natural pattern of health resource use. In these cases estimates from RCTs may have to be disregarded, or adjusted if possible.

Some resource use variables, such as treatments for rare or latent adverse events (e.g., adult-onset leukemia caused by radiotherapy in childhood), may not be available from either short-term prospective or retrospective data sources. Health economic researchers must therefore sometimes rely on expert opinion. In these cases unbiased methods using structured questionnaires should be implemented. Soliciting information from multiple experts or using a formal Delphi panel also helps to mitigate bias, although the data cannot be assumed to be accurate if they are not empirical.

Typically price variables can be obtained from existing sources. Public health plan fee schedules can be used to cost physician services for the types of procedures and assessments relevant for the analysis. In Ontario, Canada, for example, fees can be easily obtained from the online Ontario Health Insurance (OHIP) schedule of Benefits and Fees. Laboratory procedure and technical fees may also be obtained from published fee schedules. Wage, tariff, and fee schedules are also often available for other professional groups such as chiropractors, complementary practitioners, and physical therapists. Institutions may provide salary information for nurses and imaging technicians. While institutions may agree to release these values for the analysis, compensation rules may require that the actual values be suppressed in descriptions of methods and technical reports. Many public drug plan formularies provide current prices for outpatient prescription medications. The prices of prescription drugs listed on the Ontario Drug Benefit Program formulary are readily available. As described above, assigning a price to an inpatient stay is a complex task. The price varies as a function of case complexity, length of stay, and the types of cost items that are bundled into the case cost. High-quality case costs for the most responsible diagnosis associated with an inpatient admission are increasingly available as more and more institutions ascribe to management information systems that follow a uniform case costing protocol. An average bundled inpatient stay cost (i.e., case cost) based on an average length of stay can be retrieved for patients in Ontario for any given diagnosis who are admitted to one of the hospitals participating in the Ontario Case Costing Initiative. Similarly to resource use data, not all prices can be found retrospectively. Patients often pay out-of-pocket for drug dispensing, medical devices, home care, or other costs. If the items used can be identified, then it may be possible to consult wholesale price lists. Alternatively, one might include questions about out-of-pocket expenditures in prospective data collection instruments.

In collecting and assigning prices to cost items, one must be aware of the difference between prices and charges. The best value for prices is the one that represents the true value, or opportunity cost of the good or service, before any mark-ups or taxes are added. In a publicly financed, not-for-profit health care system, wage tariffs for physicians and regulated prices (before mark-ups) for medications represent reasonable approximations of the true costs. In the United States, hospital charges for specific types of admissions may be easily obtained, but are not an accurate
representation of the monetary value of the admission, since they include a profit mark-up.

Omitting indirect or time costs from an economic evaluation can bias the result if these costs are significant, such as in the case of parents caring for ill children. Time losses can also be difficult to recall, particularly for many of the items listed in Table 9.4. It is preferable to collect these data prospectively from study participants, with repeated assessments as needed. When this is not possible, sometimes simplifying assumptions are applied. For example for children admitted to hospital, one may assume that a parent will incur one day of time loss for every day of inpatient care. Just as prices represent the value of goods and services, in the human capital approach, wages represent the value of time.13 Some studies include demographics surveys that ask respondents about their earned income, usually by asking them to select from a list of ranges. These data are sensitive, and respondents may not be willing to divulge this information. Unless the respondents are a balanced representation of the target population for the study, then using respondents’ reported salaries to value time losses will limit the generalizability of the findings. Many studies are subject to volunteer bias, with study respondents often representing a more educated and higher income bracket. Using their wages to value time losses may affect the validity of the findings if time loss contributes significantly to total cost. An alternative approach is to value the reported time losses by average, national statistical wages. These data, stratified by age, sex, and occupation, may be available from national census data sources.

Cost Valuation

Once all sources for prices and resource use volumes are identified, these are listed in a costing table. Table 9.5 is an example of a typical costing table used in economic evaluations.

When stochastic data have been collected for each study patient i, cost valuation begins at the level of the individual. The cost of each item is valued by multiplying price (p) by patient-level resource use, or quantity (q). For indirect costs, the unit price is represented by an hourly or daily wage and the quantity of resource use by the patient and caregiver’s hours or days lost. Once the cost of all items is valued, a vector of total patient costs can be determined as follows:

\[ j = n \]

\[ \text{Total Cost}_i = \sum_{j=1}^{n} p_j \times q_{ij} \]

The total cost for patient i equals the sum of price multiplied by quantity for all cost items j, where the range of j extends from 1 through n cost items. With a total cost per patient, descriptive statistics can be determined including mean and standard deviation of the total costs for each treatment group. Additional statistics, such as the median, interquartile range, and minimum and maximum are also useful, particularly when cost data are not normally distributed. Differences in the mean cost per patient between groups can be tested for statistical significance using an inferential test that is appropriate for the data.

In addition to descriptive statistics on total patient level costs for each study group, it may be useful to compare the mean costs per patient for each major cost item (Table 9.6). These means can also be compared statistically, adjusting as needed for multiple comparisons. Another useful descriptive statistic is the percent contribution of each major cost items to total costs per patient. The distributions for each treatment group can be compared statistically and the results can be displayed graphically.

The above discussion of cost valuation assumes that stochastic data are available for individual study subjects. For many economic evaluations this is not the case, and instead costing proceeds based on a hypothetical well described, homogeneous cohort of patients that represents the target population. Cost item identification proceeds in the same way as described above. For cost item measurement and valuation, rather than determining health resource use at the patient level, probabilities of use of each of the cost items are determined for strata that represent different outcomes, such as the subgroups of patients that represent different levels of disease severity (and
Study Perspective

Not all possible cost items are relevant for all analyses. The cost items to be included are determined by the viewpoint or perspective of the analysis. The choice of analytic perspective hence consume different volumes of resource use) or subgroups of patients who experience different adverse events. The valuation for a cohort-based approach is described in further detail later in the cost-effectiveness analysis section of this chapter.

Table 9.5 Example of costing table

<table>
<thead>
<tr>
<th>Cost item</th>
<th>Category</th>
<th>Resource use source</th>
<th>Unit price</th>
<th>Unit price source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologic equipment</td>
<td>Intervention</td>
<td>Hospital records</td>
<td>$60,000</td>
<td>Manufacturer</td>
</tr>
<tr>
<td>Equipment depreciation</td>
<td>Intervention</td>
<td></td>
<td>$2,000 per year</td>
<td>Wholesaler website</td>
</tr>
<tr>
<td>Overhead and maintenance contracts</td>
<td>Intervention</td>
<td>Hospital records</td>
<td>$5,000 per year</td>
<td>Manufacturer</td>
</tr>
<tr>
<td>Contrast media (i.e., gadolinium injections for contrast-enhanced MRI)</td>
<td>Intervention</td>
<td>Alfa et al. 2001*</td>
<td>$37 per use</td>
<td>Wholesaler website</td>
</tr>
<tr>
<td>Digital image storage and retrieval systems (PACS)</td>
<td>Intervention</td>
<td>Hospital records</td>
<td>$35 per use</td>
<td>Hospital finance dept</td>
</tr>
<tr>
<td>Inpatient ward</td>
<td>Inpatient care</td>
<td>Hospital records</td>
<td>$450 per day</td>
<td>Hospital finance dept</td>
</tr>
<tr>
<td>Hospital overhead costs (laundry, housekeeping, utilities, capital)</td>
<td>Inpatient care</td>
<td>Hospital records</td>
<td>$130 per day</td>
<td>Hospital finance dept</td>
</tr>
<tr>
<td>Nursing</td>
<td>Inpatient care</td>
<td>Hospital records</td>
<td>$200 per day</td>
<td>Hospital human resources</td>
</tr>
<tr>
<td>Surgical procedures</td>
<td>Inpatient care</td>
<td>Bravo et al. 2002*</td>
<td>$1,750</td>
<td>Hospital records</td>
</tr>
<tr>
<td>Medications</td>
<td>Inpatient care</td>
<td>Charlie et al. 2003*</td>
<td>$400 per admission</td>
<td>Inpatient formulary</td>
</tr>
<tr>
<td>Medications</td>
<td>Outpatient care</td>
<td>Charlie et al. 2003*</td>
<td>$250 per month</td>
<td>Drug plan formulary</td>
</tr>
<tr>
<td>Home care services</td>
<td>Outpatient care</td>
<td>Delta et al. 2004*</td>
<td>$700 per visit</td>
<td>Homecare provider</td>
</tr>
<tr>
<td>Co-payment for inpatient stays</td>
<td>Direct costs</td>
<td>Resource-use questionnaire</td>
<td>$200</td>
<td>Patient report</td>
</tr>
<tr>
<td>Travel to site of health service delivery</td>
<td>Direct costs</td>
<td>Resource-use questionnaire</td>
<td>$100 per month</td>
<td>Patient report</td>
</tr>
<tr>
<td>Absence from paid labor</td>
<td>Indirect costs</td>
<td>Echo et al. 2005*</td>
<td>$1,000 per month</td>
<td>National average wage</td>
</tr>
<tr>
<td>Informal caregiver time</td>
<td>Indirect costs</td>
<td>Resource-use questionnaire</td>
<td>$750 per month</td>
<td>Patient report</td>
</tr>
<tr>
<td>Travel time to access care</td>
<td>Indirect costs</td>
<td>Resource-use questionnaire</td>
<td>$150 per month</td>
<td>Patient report</td>
</tr>
</tbody>
</table>

* Template reference, not real.
should consider the research question and the target audience—who is asking the question and who will use the value for money information generated by the economic evaluation.

As information about cost-effectiveness is used by health program decision-makers to determine whether to pay for the new service or intervention, a health care system perspective is a common approach. Other times, it is a health care institution that is the primary decision-maker. The main analytic perspectives and the relationships between them are illustrated in Fig. 9.3. Publicly financed health care programs may cover part or all of the costs of hospital services, outpatient care, laboratory tests, as well as medications for some segments of the population. Because of the breadth of services

<table>
<thead>
<tr>
<th>Table 9.6 Example table of mean costs per patient by group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard care (US alone)</strong></td>
</tr>
<tr>
<td>Physician services</td>
</tr>
<tr>
<td>Equipment</td>
</tr>
<tr>
<td>Overhead</td>
</tr>
<tr>
<td>Patient travel</td>
</tr>
<tr>
<td>Total direct costs</td>
</tr>
<tr>
<td>Indirect costs</td>
</tr>
<tr>
<td><strong>TOTAL AVERAGE COST PER PATIENT</strong></td>
</tr>
</tbody>
</table>

Abbreviations: CT, computed tomography; US, ultrasound.

**Fig. 9.3** Analytic perspectives and the relationships between them. A payer perspective can consist of separate budgets related to hospital care, outpatient services, or drug plans within the public health system; it can take a health care system perspective encompassing costs to both the public and private domains of the health care system, or it can be societal, where all costs are included regardless of payer.
covered by these plans, many economic evaluations take a public health care program perspective. However, most health care systems are a mix of private and public services. Depending on the research question, the target population, and the target decision-maker, it may be appropriate to take a wider health care system perspective that encompasses the cost items paid for by the public system as well as private third-party insurers.

In some cases the perspective may be narrow—such as with hospitals wishing to determine the cost-effectiveness of expanding their radiology services. In health care systems where the target population may be uninsured, or when patients bear a substantial portion of health care costs, then a patient or family perspective is appropriate. This would allow the capture of spillover costs incurred to other caregivers involved with the care of the patient.

Not included in any of the conventional perspectives are the indirect costs. Time losses incurred to individuals are included only when the perspective is a societal one, that is, one that considers all costs, regardless of payer. For a decision that affects a large health planning region, such as the number and location of MRI services, the principal decision-maker is the public health care program manager. While this decision-maker may be principally considered with the program’s budget and expenditures, the decision-maker must recognize that depending on the location of these services, individuals may have to travel long distances to receive care, incurring out-of-pocket costs as well as indirect productivity costs. Indirect costs may also be high for patients who are dependent on others for access to care, such as children and the elderly. Thus, although there may not exist in reality a societal payer, including a societal perspective alongside other perspectives can be highly informative and is recommended. It is not unusual to include multiple perspectives in any given analysis.

### Outcome Measurement

As seen in Fig. 9.1, the measurement of costs is weighed against the health consequences, or outcomes, experienced by the patient. The choice of analytic technique described in Table 9.1 will determine which outcome measures should be included. Table 9.7 lists some of the common outcome measures used for various interventions in radiology.

It is clear from Table 9.7 that outcomes may take a form that represents the goal of the intervention, such as cases accurately diagnosed for diagnostic interventions or improvements in health status for treatment interventions. Adverse events are also important outcomes to measure. These may include, for example, the frequency of reactions to contrast media.

<table>
<thead>
<tr>
<th>Interventions</th>
<th>Outcome measure</th>
<th>Analytic technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative imaging modalities for diagnosing appendicitis</td>
<td>Cases accurately diagnosed</td>
<td>Cost-effectiveness analysis</td>
</tr>
<tr>
<td>Insertion of intravenous access devices by interventional radiologist compared to standard operating room procedures</td>
<td>Number of complications resulting from improperly placed devices</td>
<td>Cost-effectiveness analysis</td>
</tr>
<tr>
<td>Various dosages of radiation therapy in solid tumors</td>
<td>Tumor size</td>
<td>Cost-effectiveness analysis</td>
</tr>
<tr>
<td>Various dosages of radiation therapy in solid tumors</td>
<td>Number of patients demonstrating a clinically significant response to treatment</td>
<td>Cost-effectiveness analysis</td>
</tr>
<tr>
<td>Various dosages of radiation therapy in solid tumors</td>
<td>Number of life years</td>
<td>Cost-effectiveness analysis</td>
</tr>
<tr>
<td>Various dosages of radiation therapy in solid tumors</td>
<td>Quality-adjusted life years</td>
<td>Cost-utility analysis</td>
</tr>
</tbody>
</table>
Research Methods in Radiology

or the number of adverse consequences due to exposure to radiation. It is also clear that when health improvements are studied, they may be measured and expressed as different endpoints within the same study. Endpoints such as tumor size, number of patients demonstrating a clinically significant improvement, and the number of additional life years a patient experiences are likely to be correlated. Although final outcomes are always preferred, it is not always possible to measure final outcomes such as survival or life years due to a lack of data availability, a limited time horizon, or other study design constraints. In these cases, intermediate outcomes, such as the number of patients demonstrating clinically important reductions in the size of their lesions, might be used. Intermediate laboratory endpoints, such as a normal radiograph or biomarker levels, may also be used. Intermediate outcomes can be useful when they represent the markers of health status monitored in clinical practice. When intermediate outcomes are used, however, it is predicated on the understanding that there is evidence to link the intermediate outcome to a more final, clinically relevant outcome, such as expected survival or risk of morbidity. Unless there is evidence that intermediate outcomes are significant predictors of final outcomes, they should not be used.

Quality of Life

As indicated in Table 9.7, CEAs provide the researcher with a variety of outcomes that can be used. Many diseases as well as treatments have significant impacts on a patient’s health-related quality of life (HRQoL). When quality of life effects are expected to be present, such as with radiotherapy in cancer patients, failing to measure HRQoL would be a significant limitation. There is a wide array of structured psychometrically valid HRQoL instruments available. These instruments typically consist of a series of closed-ended questions grouped in domains such as physical functioning, emotion, cognitive function, social interactions, and general health. Generic instruments such as the Short Form-36 and the Health Assessment Questionnaire are intended for use across a wide variety of diseases and conditions. Disease-specific instruments such as the European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire and the Asthma Quality of Life Questionnaire are intended for use in specific patient populations. While generic instruments are useful for comparing quality-of-life effects across diverse patient populations, they may be too blunt to detect small but clinically important changes in function. Incorporating both a generic and a disease-specific quality of life measure into a prospective study is often a useful strategy.

When HRQoL is expected to be an important outcome in a prospective economic evaluation, it is ideal to measure it with a valid preference-based instrument. Rather than simply generating a score that rates one’s HRQoL, a preference-based instrument elicits an individual’s utility, or preference, for a given health state, based on the classic conceptual domains of HRQoL. These utilities are then applied as weights to a patient’s life expectancy for the calculation of QALYs. As described above, QALYs are a universal and, therefore, powerful outcome measure. Because differences in QALYs between groups will be observed as long as the quality of life effects are present, QALYs can be calculated even when life expectancy is not expected to change. There are direct and indirect measures of preference-based quality of life. Direct measures consist of exercises that ask respondents to successively choose between health state options, allowing utility weights to be deduced. These direct approaches include the standard gamble and the time trade-off technique. Indirect measures involve administration of an instrument, such as the Health Utilities Index or the EQ-5D. These instruments incorporate a weighting algorithm in the scoring function. Direct approaches are often favored as having a stronger theoretical foundation. However, these approaches can be time-consuming and cognitively demanding compared to administration of a brief questionnaire. It should also be noted that utility is extremely difficult to ascertain in certain vulnerable populations, such as young children and individuals with cognitive impairment. The researcher must rely on a proxy reporter in these cases, such as a parent or spouse, who may have an imperfect perception of the individual’s preferences.
Time Horizon

In designing an economic evaluation, in addition to determining which analytic perspective and which analytic approach to use, one must also specify the duration of time for measurement of costs and health consequences. Establishing the time horizon for an economic evaluation should consider the interventions of interest and the characteristics of the disease being diagnosed or treated. Use of radiography to diagnose fractures may require only short-term time horizons to capture all relevant cost and health consequences. In contrast, use of imaging to diagnose and record improvements in chronic, debilitating, or life-threatening conditions such as cancer or hematological or neuromuscular disorders may require much longer, and even lifelong, time horizons. The time horizon must also be long enough to capture clinically significant adverse events that may be latent or that may occur only after years of treatment, such as those related to radiation exposure.

For many interventions costs are incurred up front or accrued in a fairly even stream over time. In contrast, health improvements are often delayed well after the start of therapy. When the time horizon of a study exceeds more than one year, the mismatch between the timing of costs and health improvements can be problematic. It is also known that while patients prefer to experience health benefits immediately, they also prefer to put off paying for it. This is known as having a positive time preference for health benefits. For this reason, analysts must distinguish between costs and outcomes that occur today versus those that are deferred into the future. This is handled by applying a discount rate to costs and outcomes when they are measured over a year or more. The same constant rate should be applied to both costs and health consequences and this rate is typically 3% to 5%, reflecting society’s rate of positive time preference.4

It becomes clear that when long time horizons are needed, the ability to collect all necessary data becomes challenging and often rules out a prospective study design. For this reason, the use of decision analysis to model the probabilities of various intervention and treatment pathways and to model the costs and outcomes associated with these pathways occurring in a hypothetical patient cohort can be useful. This is explained in further detail in Chapter 8.

Cost-Effectiveness Analysis

As described at the start of this chapter, a cost-effectiveness analysis is a full economic evaluation whereby both costs and consequences of health care programs or treatments are examined, the health consequences are measured in natural units (e.g., life years), and the outcomes measured are common to all comparators. The results of a CEA are expressed as the incremental cost per unit of effect, such as, dollars per life year gained.

The unit of analysis for costs and outcomes is often the patient. When the measure of effectiveness is a proportion, such as the percent of patients achieving a clinically significant response, it can be easier to interpret the analysis if the costs and outcomes are scaled to a group of 100 or 1,000 patients.

The point estimates of the means for each treatment group are represented in an incremental cost-effectiveness ratio (ICER):

\[
\frac{(\text{Cost}_I - \text{Cost}_{SC})}{(\text{Effectiveness}_I - \text{Effectiveness}_{SC})}
\]

The numerator of the ICER is the difference in mean costs between the experimental intervention (I) and standard care (SC) (or another comparator). The denominator expresses the difference in mean outcomes between comparators. By subtracting the mean values for costs and outcomes of standard care from the intervention, one can determine the added or incremental costs associated with the intervention per unit of added or incremental benefit achieved.

ICERs therefore allow the direct comparison of two comparators in terms of costs and outcomes. If data on survival and utility related to treatments are available, then a cost-utility analysis that examines the incremental cost per QALY gained can be conducted. Finally, the results of the analysis are presented in a table depicting the incremental values and the ratios (Table 9.8).
Decision analysis is a common technique for undertaking an economic evaluation, particularly when stochastic patient-level data are not available for all variables. In decision analysis, input data are used to construct a decision tree based on a hypothetical cohort of patients. Decision analysis allows multiple interventions to be compared to standard care in a single tree. In this way the costs and health benefits associated with each intervention can be directly compared to each other as well as to standard practice. Decision analysis is described in more detail in Chapter 8.

A separate analysis should be conducted for each analytic perspective (societal, health care system, institution, family) whereby the total cost estimate is varied. It may also be useful to conduct separate analyses for homogeneous hypothetical subgroups of patients that are clinically distinct, such as by age group, diagnostic subtype, or presence of risk factors. The analysis inputs would then be customized for these subgroups.

A common alternative decision model that can be used, particularly for chronic conditions, is a state-transition Markov model. State-transition models can be incorporated directly into decision trees.

**Uncertainty**

It is clear that economic evaluations require large amounts of accurate data. Sometimes not all variables can be obtained from high-quality data sources. In the case of missing or uncertain data, assumptions may be needed. For example, if physician fees are not available for the jurisdiction where the work is being performed, then fee schedules from a neighboring region may be used with the assumption that the values are similar between regions. When data on the frequency of follow-up visits to a community physician over the study time horizon are not available, a panel of experts might be convened to provide estimates based on their practices or based on guidelines for recommended care. Survival estimates may not be available for children with a certain condition but might be reasonably extrapolated from adult data, adjusting for differences in actuarial life expectancy between children and adults. Assumptions may be needed for the base case estimate of a variable or for the upper and lower bounds that are used to indicate the precision of the point estimate. For every study, the researcher must state all assumptions explicitly, with a rationale for the base case value used in the analysis. Sensitivity analysis is employed to assess the sensitivity of the incremental costs, incremental effects, and the ICER to changes in the assumptions.

Sensitivity analysis includes deterministic and probabilistic approaches. In a one-way deterministic sensitivity analysis, the base case values for uncertain variables are replaced by alternative values, such as upper and lower 95% confidence intervals (CIs), one variable at a time. This allows the researcher to determine to what extent uncertainty in a single variable contributes to overall uncertainty in the ICER.

For example, in a study comparing ultrasound or CT performed in an emergency department for the detection of appendicitis, the timing of the ultrasound or CT may affect the sensitivity and specificity of either approach, since a delay in imaging may allow...
natural progression of appendicitis making it easier for the sonographer to visualize the inflamed appendix or adjacent inflammatory signs (fewer false negatives). A one-way deterministic sensitivity analysis that varies the sensitivity and specificity will enable one to investigate the effect of timing on the ICER comparing ultrasound to CT.

Sensitivity analysis results are depicted in a tornado diagram, with the variables ranked from having the most effect to having the least effect on the ICER. An example of a tornado diagram is depicted in Fig. 9.4.

In a multiway deterministic sensitivity analysis, two or more uncertain variables are varied simultaneously. This approach is helpful in illustrating the effect on the ICER of changing one uncertain value, such as admission cost, for alternative fixed values of a second variable, such as test prices.

Advanced probabilistic techniques can be used to directly incorporate all the uncertainty into the analysis when a base case and a range are specified for each variable in a decision analysis. In a probabilistic sensitivity analysis, the base case estimate and a range, often based on 95% CIs or standard deviations, are specified for every variable, along with a distribution (e.g., normal, beta, gamma, log-normal, uniform). Through Monte Carlo simulation, the data are sampled 1,000 times or more. Each iteration produces values for each variable, incremental costs, incremental benefits, and an ICER. From the 1,000 samplings of the data, a probability that the intervention is cost-effective can be deduced. This is essentially the proportion of ICERs that fall in the southeast quadrant. A scatter plot (Fig. 9.5) is a useful visual depiction of the results of a probabilistic sensitivity analysis. By examining the proportion of iterations that fall within each of the quadrants, one can infer the probability that the intervention is dominant (proportion of points in southeast quadrant).

Because of the inherent uncertainty in economic evaluation, the results of sensitivity analyses are as important as presenting the base case ICER. Sensitivity analyses convey
Communicating the results should take the form that is most advantageous to the target decision-maker. This may be a briefing note and an executive summary with limited technical jargon. Often the decision-maker wishes to determine whether a new intervention is cost-effective—that is, make a value-for-money decision, under varying thresholds for willingness-to-pay. A decision-maker may have a higher or lower threshold for willingness-to-pay for any given intervention depending on the jurisdiction, societal values regarding the health gains and who is incurring them, and budget constraints.

To determine the net monetary benefit of an intervention, a decision-maker’s willingness-to-pay threshold, say $50,000 for a QALY gain, additional important evidence that allows the target user to judge their confidence in the allocation decision that flows from the findings.

**Synthesis, Reporting, and Knowledge Translation**

The job of a health economist does not end with the completion of the economic evaluation. The final step is to ensure that the evidence gets into the hands of the decision-makers who need it. Ideally these individuals, representing clinical, institutional, or jurisdictional decision-makers, will be involved as a partner from the first step of articulating the research question. Reporting back closes the loop to provide the necessary evidence.

Communicating the results should take the form that is most advantageous to the target decision-maker. This may be a briefing note and an executive summary with limited technical jargon. Often the decision-maker wishes to determine whether a new intervention is cost-effective—that is, make a value-for-money decision, under varying thresholds for willingness-to-pay. A decision-maker may have a higher or lower threshold for willingness-to-pay for any given intervention depending on the jurisdiction, societal values regarding the health gains and who is incurring them, and budget constraints.

To determine the net monetary benefit of an intervention, a decision-maker’s willingness-to-pay threshold, say $50,000 for a QALY gain, additional important evidence that allows the target user to judge their confidence in the allocation decision that flows from the findings.
is assigned. If a new intervention yields QALY gains that when converted to dollars with the selected threshold yield more benefits than costs, then the intervention is deemed cost-effective. If the willingness-to-pay for a QALY gain is very high, then there is a high probability than the monetary health benefits will exceed the costs. This can be illustrated in a cost-effectiveness acceptability curve that plots the probability of cost-effectiveness (incremental benefits exceed incremental costs) against decision-maker thresholds for willingness-to-pay. An example of a cost-effectiveness acceptability curve is depicted in Fig. 9.6. In this example there is a 60% probability of the intervention being cost-effective at a decision threshold of $50,000 per QALY gained.

A full formal technical report with technical appendices that describe advanced methods and provide additional analyses and detailed cost calculations should always be made available. It can be posted to the researcher web page for download and should also be included in international repositories of health technology assessment, such as the York Centre for Reviews and Dissemination. The research team should actively seek out opportunities to present the findings to multiple stakeholder audiences. The plan for report writing and presentation can be laid out in a knowledge translation plan that is part of the original study proposal. This may include disseminating a report summary to key clinical organizations, agencies, and institutions. Researchers will also naturally want to publish their findings in a peer-reviewed medical or health policy journal where it will reach a wider audience, as well as present the findings at a scientific conference.

In all reports, publications, and presentations, the researcher must demonstrate transparency with regard to all methodologic approaches and assumptions. An honest reckoning of study limitations must be provided. The issue of generalizability to other patient groups, other institutions or other jurisdictions must also be addressed, particularly if the report is widely disseminated. The researchers must also be up front regarding all funding sources and potential conflicts of interest. Any plans to update the analysis, for example, in the event that a new generation of technology is introduced, should be described. This is relevant for radiological imaging technologies

**Fig. 9.6** Example of cost-effectiveness acceptability curve. In a cost-effectiveness acceptability curve, the probability that an intervention is cost-effective (incremental benefits exceed incremental costs) is plotted against a decision-maker’s willingness-to-pay budgetary threshold. In this example, there is a 60% probability that the intervention is cost-effective at a decision threshold of $50,000 per quality-adjusted life years (QALY) gained.
that evolve rapidly. Often the model can be retained with some or all inputs updated.

In summary, in a society with a complex health care system characterized by practitioners that aspire toward excellence, health care consumers expect and demand the very best in health care. This includes timely access to state-of-the-art treatments and services. At the same time, the North American population is aging and consuming more health care while growing economic pressures put constraints on health care spending. An increasing demand for health services comes not only from consumers. Our health care system also functions much like a market place. The “technological imperative” to continually innovate has spurred the biotechnology sector to expand the use of technology in health care delivery, notably in the field of radiology. Thus, suppliers also exert pressure on the system, exhorting practitioners to stay on the “cutting edge” of care while pressuring payers to invest in rapidly evolving medical technology.

In the widening gap between the demand for high quality health care and the availability of limited resources to deliver that care, opportunistic inequities can arise. The answer is not to spend more on health care services, but to increase efficiency and consume only the resources that have demonstrated health benefits. Health economic evaluation helps to reduce inequity and inefficiency by producing evidence that allows decision-makers to make allocation decisions that maximize health benefits for dollars spent. Like clean air and water, health care is a precious resource that requires our utmost attention. This must come in the form of high-quality evidence for health care allocation decision making.

■ Highlights of Key Points

• The goal of health economic evaluation is to determine which intervention, program, or health service option is the one that is most economically efficient—that is, that maximizes health gains per dollar spent and achieves value for money.

• The four types of economic evaluation are cost-effectiveness analysis (CEA), cost-utility analysis (CUA), cost–benefit analysis (CBA), and cost–minimization analysis (CMA).

• The research question should flow from an identifiable problem and should address a gap in existing evidence. It should be posed in terms of costs, outcomes, comparators, perspective, and target population.

• The costing process is comprised of cost item identification, cost item measurement, and valuation.

• The cost items that are included in an economic evaluation can be broadly categorized as direct health care–related costs, direct patient (out-of-pocket) costs, and indirect costs.

• Many diseases as well as treatments have significant impacts on a patient’s HRQoL, and these should be measured using a suitable preference-based instrument.

• Analysts must distinguish between costs and outcomes that occur today versus those that are deferred into the future. This is handled by applying the same constant discount rate to costs and outcomes when they are measured over a year or more.

• The results of an economic evaluation can be reported as an ICER, with the numerator representing the difference in mean costs between the experimental intervention and standard care (or another comparator) and the denominator representing the difference in mean outcomes between comparators.

• A separate analysis should be created for each analytic perspective (societal, health care system, institution, family) whereby the total cost estimate is varied. It may also be useful to construct separate analyses for homogeneous hypothetical subgroups of patients that are clinically distinct, such as by age group, diagnostic subtype, or presence of risk factors.

• Sensitivity analysis includes deterministic and probabilistic approaches. In a one-way deterministic sensitivity analysis, the base case values for uncertain variables are replaced by alternative values, such as upper and lower 95% CIs, one variable at a time.
In a probabilistic sensitivity analysis, the base case estimate and a range, often based on 95% CIs or standard deviations, are specified for every variable, along with a distribution (e.g., normal, beta, gamma, log-normal, uniform). Through Monte Carlo simulation, the data are sampled multiple times. Each iteration produces values for incremental costs, incremental benefits, and an ICER.

To determine the net monetary benefit of an intervention, a willingness-to-pay threshold, say $50,000 for a QALY gain, is assigned. If a new intervention yields QALY gains that when converted to dollars with the selected threshold yield more benefits than costs, then the intervention is deemed cost-effective.

References
4. Guidelines for the economic evaluation of health technologies: Canada. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2006
Learning Objectives

- To provide the basis for conducting research and publishing results in scientific journals outlining the steps for successful accomplishment of tasks related to formulation of research question(s), search of up-to-date literature related to the topic of investigation, conduct of research, drafting of manuscripts, review process by authors, submission of manuscripts for consideration of publication, and reply to journal’s editors and reviewers’ comments.
- To discuss frequently asked questions concerning authorship.

Introduction

Reasons for Publishing

The evidence found in peer-reviewed publications constitutes the foundation for advances in clinical practice. Conducting and disseminating clinically relevant research upon high ethical and editorial standards is imperative for improvement of patient care. Further, publication may enable advancement of individual career goals. Although the portfolio of the academic radiologist encompasses clinical, research, and teaching activities, the dogmatic paradigm “Publish or Perish” prevails for a successful academic career considering the fact that publication is a metric parameter of scholarly productivity in association with other parameters such as successful grants and awards. The assessment of an individual’s publication track record is an important part of the academic promotion and tenure process.

In this chapter we discuss the steps for preparation of a manuscript as the final output of a research project from organizing tasks in a timely fashion to accruing information obtained throughout the research conduct, finalizing the manuscript draft, and submitting it for consideration of publication in a scientific journal.

Steps of Research Conduct

Preparation for the Conduct of a Research Project

Previous studies have shown that the chances of professional success in a given activity substantially increase with higher levels of “preparation.” In addition, authors have been attributed tasks that are not urgent, but important, such as planning and preparation (writing proposals, manuscripts, and grants), studying, prevention of conflicts, and relationship building, as the key components that efficient workers apply on a day-by-day to achieve professional success.

Trainee as the Primary Author

Before the trainee contacts potential supervisors with respect to conducting a research project he/she should be able to juggle his/her available time between clinical and educational duties during his/her internship, residency, or fellowship. This includes devotion of time where he/she can perform at his/her best avoiding pre- and post-call periods that can be stressful and exhausting. Booking academic days in the beginning of the week (Monday or Tuesday) enables trainees to catch up and finalize pending tasks such as communicating with supervisors and collaborators, submitting documents, and completing short activities during the week.

It is essential that the trainee only agrees with the conduct of research projects that he/she will be able to finalize within the proposed...
Conducting and Publishing Research and collaborators with a strong track record of publications and research grants, if available. The trainee should aim to find a supervisor who has a proven track record of coming up with project ideas and successfully mentoring other trainees through the process from start to final publication. Selecting a supervisor who practices in an institution geographically close to the locale of the trainee’s rotations may facilitate (but is not exclusive) trainee–supervisor communication along the project. If the trainee does not have his/her own ideas for a research project, then he/she can inquire for ideas from his/her supervisor. Having a supervisor for the research project does not exclude the possibility that the trainee looks for further guidance from a co-supervisor who typically has expertise in an area of relevance for the research project (e.g., epidemiology, rheumatology, orthopaedic surgery) different from the area of expertise of the supervisor (e.g., musculoskeletal radiology).

The trainee should investigate the feasibility of the project before committing to it. Is there a sufficient number of patients or imaging studies meeting criteria for the study? Can they be adequately accessed? Does the study have adequate power? Is there a good reference standard? Is the allotted time to complete the study appropriate? Many studies initiated by trainees and young faculty fail because they weren’t feasible to start with. A good supervisor should aid the trainee to identify what questions to ask about feasibility and to make these determinations.

Co-Investigators

There is not a predetermined maximum number of investigators for a research project; however, each investigator should have a well-defined role in the project that should be determined prior to the commencement of the study. Depending on the level of difficulty of the research project, more investigators may be required such as in the case of multicentric clinical trials where investigators from different institutions are committed to participating in different steps of the conduct of the research project.
Definition of the Research Question(s)

The research question should be based on the PICO (P, Patient/Problem; I, Intervention; C, Comparator/Control; O, Outcome) concept which was developed for formulation of questions and searching for evidence as discussed in Chapter 1. Nevertheless, prior to finalizing this step the investigators should ensure that the overarching research question(s) can add relevant information to the available bulk of knowledge in the literature. The candidate study can be complementary to prior research in a given field, but ideally should not propose research questions similar to those already proposed in previously published studies, unless there is a reason for this, such as the existence of controversy on the results of previous studies on a given topic. In this case the current study should hold potential for providing better evidence on the topic of interest (e.g., sample size with sufficient power, more rigorous methodology, multicentric characteristics, etc.).

Conduct of a Power Literature Search

To accomplish a broad review of the existent literature on a given topic, a power literature search is recommended. This search has two objectives: (1) to determine the legitimacy of the proposed research question within the context under investigation, that is, to determine if the project is sufficiently novel; and (2) to provide a summary of the available publications in the literature that have common search term(s). It will help the investigators to frame the current project by providing a knowledge base of what is known and what requires further investigation, thus helping the authors to design a better study.

The steps for a power literature search are: (1) to create a searchable question using a concept map or PICO, (2) to select an appropriate resource such as a database, (3) to run a search using database subject headings, or text words, (4) to combine results for each search term using Boolean operators, (5) to review search results, (6) to repeat steps 3 and 4 with additional terms if search results are not relevant, (7) to save search results, (8) to save search strategy (search steps), and if desired, (9) to create auto alerts.

Creation of a Searchable Question Using a Concept Map or PICO

To find articles, books, or other materials on a subject, the investigator will need to develop a search query from the research question. This involves breaking down the research question into main ideas. There are different methods that can be used to help with the process of identifying the main ideas or key concepts.

One method is to create a concept map. For this purpose the investigator should: (1) write the question in sentence form—it can be the “perfect article title”—something reflecting exactly what is being sought; and (2) underline the key concepts. For example: Is MRI more effective than radiography (X-rays) in diagnosing rheumatoid arthritis? Each underlined concept can be placed in its own column in a table. Within each column of the table, the investigator should add synonyms for the underlined concept, alternative spellings, and alternative word endings. In a search for literature, the terms within a column are combined using the Boolean operator “or”; the terms in separate columns are combined using the Boolean operator “and.”

The PICO method is another way of breaking down questions into smaller concepts. A PICO for the search question “Is MRI more effective than radiography (X-rays) in diagnosing rheumatoid arthritis?” would be:

- **Patient/Problem:** Rheumatoid arthritis
- **Intervention:** Radiography
- **Comparison:** MRI
- **Outcome:** Effectiveness

Selection of a Database

Once the search concepts have been identified, the next step is to select appropriate resources within which to conduct searches. Databases are one example of a resource that can be used to identify relevant literature. They are collections of information that
Use of Subject Headings, or Text Words

Many databases like MEDLINE and EMBASE have their own unique search vocabulary. This controlled vocabulary, also called subject headings, consists of standardized terms used to tag, or describe, the contents of an article. Subject headings are assigned by indexers to reflect what is discussed in the article. It is important to become familiar with, and use, the appropriate controlled vocabulary or database subject headings when conducting searches. The system of medical subject headings used in MEDLINE is called MeSH for short. The underlined terms in a concept map, or the PICO terms, can be used to help identify pertinent database subject headings.

In the cases where the search topic is a rare syndrome, new technology, new drug, new catchword, or brand name, it is very likely that there will not be a relevant database subject heading. The researcher will need to conduct a text word search. In this type of search, terms entered into the search interface are matched against word(s) that appear in specific fields such as the title and/or abstract of an article. To ensure that articles are not missed when conducting a text word search, it is important to include synonymous terms, singular and plural forms of words, as well as variations in British and American spelling. The Help section of the search interface for a database provides information about features such as truncation, wildcards, proximity or adjacency commands, and field searching that can be used to create efficient, complete text word searches.

Literature searching involves balancing specificity and sensitivity. Specific searches retrieve a higher ratio of relevant to irrelevant references. For example, searches conducted using database subject headings tend to retrieve more references that are relevant. Sensitive searches are designed to find all possible references and, as a result, retrieve many that are irrelevant. Text word searches tend to retrieve more irrelevant references; search term(s) may be present in the title or abstract, but not used in the sense that is required. An example of a specific search would be a literature review search—just key articles are needed. Systematic review searches, on the other hand, need to be sensitive. They should incorporate both database subject headings and text words to reduce the probability of missing any potential relevant records.

Once the type of search needed—specific or sensitive—has been determined, the next step is to begin searching. Rather than typing all the search terms into the database interface at once, it is preferable to look for each concept separately. For example, search for all the references about rheumatoid arthritis, then look for all the references about radiography, and finally all the references about MRI. In most database search interfaces, the results retrieved with each search concept are assigned a unique set number. These set numbers can be used to combine the different results together using the appropriate Boolean operators. When reviewing the final set of results, be alert for additional terms that may be useful. These additional terms can be searched and combined with the sets of previous terms.

Combination of Results of Search Terms, Review and Saving of Results, Repetition of Steps, Creation of Auto Alerts

All the sets created during the search process form what is referred to as the search strategy. It is worthwhile taking a few moments at the end of a search to do two things: save the search strategy and create automated updates. Saving the search strategy means saving all the steps in the search. The saved search can simply be run again at a later time rather than having to re-create the strategy from scratch. In addition, it may be necessary to provide a brief description of the search in the Methods section of the manuscript and to include the search history as an appendix. Having a document of the search strategy makes this process easier. As it is likely that there will be a long period of time between the initial literature search and the preparation of a final
Research Methods in Radiology

manuscript reporting the results of a project, it will be necessary to track new publications on the topic subsequent to the initial search. The saved search strategy can become the basis of an alert about these new publications. Results of the searches may be emailed or saved in a user-created account. Depending on the search interface, an alert can be set to run daily, weekly, monthly, or quarterly. The name for an alert varies by interface: Auto Alerts (Ovid interface), Email Alerts (PubMed interface), or Alerts (Web of Knowledge interface).

Investigators are encouraged to use a reference management software program (e.g., Zotero and Mendeley, are freely available on the web) to store, in a library or database, information about relevant articles on the topic of study. These can be the results of the searches imported directly from the database interface into the reference management program. Many of the reference management products can link with popular word-processing programs to let authors insert citations from the database or library into a document as they are writing. The software will format the citations and create a reference list or bibliography in accordance with the specific instructions of a selected journal or publisher. During the process of document revision, the citations and reference list or bibliography will be updated as changes are made.

Design, Methodology, and Sample Size

Discussions with a Statistician

Once Sir Ronald Fisher (February 17, 1890 to July 29, 1962), an evolutionary biologist, geneticist, and statistician, said: “To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.”

Statisticians can assist researchers in designing the research question(s), selecting the study design, defining appropriate outcomes, and offering suggestions on design implementation. As part of the feasibility aspect of a research project, it is extremely important for the investigators to estimate the required sample size (Chapter 15). It is also recommended (and solicited by most institutional Research Ethics Board committees) for the investigators to prepare a database spreadsheet for the project prior to its commencement, which should be discussed with the statistician ahead of time. Examples of databases for radiology projects are available in Chapter 11.

Authorship and Contributorship

Discussions among Investigators

It is recommended that discussions about authorship should be initiated by the principal investigator of the project (which can be the trainee’s supervisor) early during the conduct of the project. Several guidelines are available on the web to determine minimum standards for authorship of manuscripts.8–11 A priori agreement among the project’s investigators on the adherence to these guidelines may reduce the possibility of future disagreements on authorship among investigators at the phase of finalization of the project.

An author is considered someone who has made relevant contributions to the bulk of knowledge of the publication, and who continues to hold important academic and financial commitments with the published work.12 As per the International Committee of Medical Journal Editors (ICMJE) an author must take responsibility for at least one component of the work, should be able to identify who is responsible for each other component, and should ideally be confident in their co-authors’ ability and integrity.13 In the past, readers were rarely provided with information about contributions to studies from persons listed as authors and in acknowledgments.14 Nevertheless, some journals now request and publish information about the contributions of each author of the study, at least for original research.

Although contributorship and guarantorship policies remove much of the ambiguity surrounding contributions, they are unable to solve the question of the quantity and quality of contribution that qualify for authorship. Because of this, the ICMJE has determined three minimum requirements for authorship credit on manuscripts submitted to medical journals since 199713: (1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; (2) drafting
voluntarily excludes him/herself from authorship on request; otherwise it is considered misconduct or poor ethical conduct. A known example of misconduct ghost authorship relates to the case of Rosalind Franklin, who was a Cambridge-educated chemist who took the first X-ray diffraction picture of DNA (Photograph 51). She showed the Photograph 51 picture to her colleague and co-author Gosling, who without Franklin’s knowledge or permission, showed it to a scientist named Wilkins, a competitor of Franklin’s. Wilkins showed it to Watkins, who looked at it and immediately realized that DNA was helical, which led directly to his theory. He then drafted and published a paper that did not acknowledge the photograph directly. In 1953, *Nature* published three papers: (1) one by Watson and Crick; (2) one by Wilkins, the collaborator who had shown Photograph 51 to Watson; and (3) one by Franklin and Gosling, which described Photograph 51. Watson did not cite Franklin’s paper in his article, only acknowledged it in an endnote at the bottom.

In 1958 Franklin died of ovarian cancer at the age of 37, and was not eligible to receive the Nobel Prize posthumously. In 1962, Watson, Crick, and Wilkins received the Nobel Prize in Physiology or Medicine for their proposal on the DNA double helix structure. The fact that Franklin took the picture but could not recognize what it meant at that moment does not take the credit from her work. This case illustrates the need for proper scientific credit and recognition of collaborative work.

### Documentation of the Project

The documentation of a research project has three major parts: (1) description (Materials/Patients and Methods); testing (Results); and analysis of results (Discussion) (Table 10.1). It can be obtained during and/or after the conduct of the project. Although there is no limit for the maximum number of authors who contribute for the drafting of a manuscript, it is advisable that a single senior author carefully reviews the entire document to ensure that the document has a logical sequence and coherence, and that different parts of the document connect with each other.
### Table 10.1 Details on manuscript sections

<table>
<thead>
<tr>
<th>Methods subsections</th>
<th>Question to be answered</th>
<th>Type of study population</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>“Who were the target groups?”</td>
<td>Humans</td>
<td>Reporting on number of subjects, age, sex, severity of disease, region of interest, patient recruitment, inclusion/exclusion criteria, study model (prospective or retrospective), date for start and end of the study, databases used, reasons for failure for testing some subjects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Animals</td>
<td>Reporting on animal model, species, baseline weight/age, sex, region of interest, animal randomization (if applicable), inclusion/exclusion criteria, pathology method, storage requirement for postmortem tissues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cadavers</td>
<td>Reporting on number of specimens, age, sex, region of interest, cause of death, pathology method, storage requirement for postmortem tissues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computer model</td>
<td>Features that were the base for the development of the model and the rationale for the model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Systematic reviews/ meta-analysis</td>
<td>Protocol and registration, eligibility criteria of study subjects, databases and other sources used, primary studies’ search and selection</td>
</tr>
</tbody>
</table>

| **Testing**          | “How were the target groups tested and the outcome data collected?” | Humans                   | Reporting on administration of databases, tests and reference standard, on how (techniques for blinding or unblinding of operators during scanning and selection of regions of interest during data interpretation, time interval between tests and reference standard), by whom (number, training, and experience of operators/readers), and when outcome measures were obtained, on techniques (parameters for each imaging modality) and scanners (brand, model and city/state of scanners’ manufacturers, software, etc.) used for data acquisition and analysis, time points for data collection, primary and secondary outcome measures, adverse events |
|                      |                          | Animals                   | Reporting on scanners used, the parameters applied/evaluated, time points for data collection, primary and secondary outcome measures, adverse events |
|                      |                          | Cadavers                  | Same as for animal studies |
|                      |                          | Computer model            | Reporting on type and value of input parameters used, software information (brand, city/state of inventors, version year), “trigger criteria” to determine solutions |
|                      |                          | Systematic reviews/ meta-analysis | Description of methods used for assessing risk of bias (Cochrane system) or quality of reporting (STARD) or diagnostic accuracy methodology (QUADAS-2) or quality of evidence (GRADE) of primary studies, how the assessment of risk of bias may affect cumulative evidence, summary measures |

(continued)
Prior to starting the write up of the manuscript, the authors should define the manuscript type (original research, technical development, review, editorial, letter to the editor, survey, case report, pictorial essay, etc.) as per guidelines of the scientific journal to which the manuscript will be submitted. The authors should also review the Information for Authors document of the journal (including maximum number of words format for the text, tables, and figures). In this chapter, we focus on the process for preparation of original research manuscripts. Investigators should read, understand, and adhere to the journal’s Information for Authors (Author Instructions). Lack of adherence to the journal’s instructions raises concern from the editors and reviewers for how careful the investigators have been with their project conduct.

The sections of the manuscript should be organized as shown in Table 10.2.

In the following paragraphs, we describe a suggested chronologic order for preparation of manuscript draft sections.

### Materials/Patients and Methods Section

The Materials/Patients and Methods section of a manuscript allows readers to verify the methodology used in the study, which has implications, along with other factors, on the quality of the study. If the project has been well thought out in advance and has a well-written proposal, part of this section should already be available in the proposal, which is a good starting point for the first manuscript draft. Ideally the Materials/Patients and Methods section is written before the project is conducted, not after. It is advised that the first sentence of this section states whether institutional Research Ethics Board approval has been received for conduct of the study and whether informed consent has been obtained from patients or patients’ parents, if applicable, for clinical studies. Likewise, for animal studies, authors should state whether the study protocol was approved by the institutional animal care committee and whether it complied with the local Council (or equivalent organization) on Animal Care Guidelines.

The information available in the Materials/Patients and Methods section of a manuscript should be described in sufficient detail (references to methods used, including statistical methods, and rationale for using new or modified methods) to allow other investigators to reproduce the results of the study.
Table 10.2  Outline of common sections to be included in manuscript submissions

<table>
<thead>
<tr>
<th>Order of files</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cover letter</td>
<td>Cover letter to the journal’s editor</td>
</tr>
<tr>
<td>2. Title page</td>
<td>Abbreviated title page (includes complete title of manuscript, “short running title” if requested, all authors’ names including middle initial in the appropriate order as well as corresponding author information, affiliation, mailing and email addresses, telephone and fax numbers, and source of support)</td>
</tr>
<tr>
<td>3. Manuscript body</td>
<td>Blank page (includes title and type of manuscript: original research, review, pictorial essay, etc.) Abstract (number of words), Key words Introduction Materials/Patients and Methods Results Discussion/Conclusion/Number of words (from Introduction to Conclusion) Acknowledgment References Tables</td>
</tr>
<tr>
<td>4. Figures</td>
<td>Figures (no composite images; each image should be a separate figure part and represent a separate image file; for each figure, a separate tiff file with a corresponding legend should be provided; insert legends by the time the figures are uploaded onto the journal’s website). One figure may contain multiple images, which should be labeled as a,b,c,d, etc.</td>
</tr>
<tr>
<td>5. Supplemental material (Appendix)</td>
<td>Online-only publication of supplemental material allows publication of material that cannot be accommodated in print. This material can include (a) multimedia (e.g., animation, dynamic image sets [movies], audio), (b) large numbers of relevant images whose number would exceed the limits of print publication, (c) relevant data in the form of tables or text that could not be accommodated in the print version, and (d) interactive materials such as Java applets and other programs for expanding browser capabilities and interactivity in areas such as image display and computer-assisted instruction</td>
</tr>
<tr>
<td>6. *Financial disclosure</td>
<td>Financial disclosure may be required for each author</td>
</tr>
<tr>
<td>7. *Conflicts of interest statement</td>
<td>Conflicts of interest statement may be required for each author</td>
</tr>
<tr>
<td>8. *Transfer of copyright agreement</td>
<td>Transfer of copyright agreement is required for each author</td>
</tr>
<tr>
<td>9. *Authors’ contributions</td>
<td>Some journals require specifications of individual contributions of authors according to the ICMJE submitted by the guarantor of the integrity of the work who can be the senior author or all authors</td>
</tr>
</tbody>
</table>

Abbreviations: ICMJE, International Committee of Medical Journal Editors.

If the reader cannot reproduce the study after reading the Materials/Patients and Methods section of a manuscript, then it is insufficient. It is recommended that authors follow the checklist of the Standards for Reporting of Diagnostic Accuracy (STARD) statement for studies dealing with diagnostic accuracy, the Consolidated Standards of Reporting Trials (CONSORT) statement for randomized controlled trials, and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) tool for meta-analyses of randomized controlled trials. The authors should avoid nonstandard terms, novel or ambiguous abbreviations, and self-aggrandizement of their own work (e.g., “excellent,” “innovative,” “unique”). The use of abbreviations should be minimized. If standard abbreviations are used in the text, the authors should define/spell them out in full when it is first used in the
abstract, in the text, and in each figure legend. Abbreviations should be defined in each table, usually in a footer.

For example, for a journal that considers 3,000 as the maximum limit of words for an Original Research manuscript, e.g., Radiology, the Materials/Patients and Methods section is expected to not exceed 800 words.21

Results Section

Once the statistical analysis of the data has been performed and the preliminary or final results become available, the investigators should discuss them among themselves and with the study statistician in order to determine the face validity of the results that relates to their ability to be credible22 (whether the results make sense either clinically or experimentally). After ensuring that the results are accurate, credible, and were obtained using appropriate statistical methods, the investigators should hypothesize rationales to justify the findings. The PI can keep a log file of the investigators’ discussions during meetings. The assigned investigator of the study can then start drafting the Results section of the manuscript. The Results section should mirror the Materials/Patients and Methods section. Corresponding results should be presented in the same order as in the methods in the Materials/Patients and Methods section. There should be nothing in the Results section that was not introduced in the Materials/Patients and Methods section.

The Results section should be short, but contain at least three paragraphs comprising: (1) demographics of the population (this information could be presented in a table format instead); (2) results of the research questions; and (3) adverse events, if applicable. The results should be presented in logical sequence in the text, along with tables and illustrations. The authors should not repeat information that has already been covered in tables and/or figures. Only important information should be summarized in the text, tables, or figures. Results should be given for all items evaluated as mentioned in Materials/Patients and Methods. Numerators and denominators must be provided either in the text or in the tables for all percentages given. P values and confidence intervals should be provided wherever appropriate.

The Results section is expected to not exceed 1,000 words for a journal that considers 3,000 as the maximum limit of words for an Original Research manuscript, e.g., Radiology.21

Discussion Section

After the results have been summarized in the text of the manuscript, and before writing the discussion, the authors should update the literature search for new papers in the field of study. If the authors have created alerts for new citations at the time of the initial search then they should select and retrieve papers that may support or contradict their study results. The authors should highlight the relevant concordant and discordant information and add this information in the Discussion section of their manuscript, as appropriate, following a sequential order. Each paragraph of the Discussion section should correspond to one paragraph of the Results section. If new citations provide information similar to that generated by the author’s study the authors should try to explain (if possible) how their study results can complement the published information. The process of drafting the Discussion section can be the most challenging and time-consuming aspect of writing a manuscript. It requires a broad knowledge of the information that is available in the literature and the ability to relate the results of the study to those reported in other relevant studies raising points of similarity and disagreement. The authors should also raise hypotheses to justify unexpected results of the study and report limitations of the study design, techniques used, and overall conduct. Setting aside sequential academic days for drafting the Discussion section can facilitate the production of a cohesive document.

The Discussion section should emphasize the advances in knowledge made by the study, which can be made in the first paragraph of the section. The authors should not repeat results in detail, but should rather include the implications of the findings and their limitations. On the final paragraphs of
this section the authors should address study limitations. A good discussion of limitations shows that the authors have the ability to critically assess their own work. If the limitations are written well, they will adequately anticipate and address concerns that reviewers identify with the study. Limitations should be written within a paragraph (or paragraphs) at the end of the discussion, just prior to the conclusion.

The Discussion section should conclude by linking the conclusions with the goals of the study, avoiding unqualified statements and conclusions not supported by the data. For articles on experimental studies, the authors should describe the importance of the conclusions as they relate to potential future practical applications. Details of the Discussion section are available in Table 10.3.23

For example, for a journal that considers 3,000 as the maximum limit of words for an Original Research manuscript, e.g., Radiology,21 the Discussion section is expected to not exceed 800 words.

Upon completion of the first draft of Materials/Patients and Methods, Results, and Discussion, the authors should re-read all the document, focus on refining thoughts, correct typos, and start drafting the Introduction and Abstract sections.

### Table 10.3 Suggestions for drafting the discussion section of a manuscript

<table>
<thead>
<tr>
<th>Discussion subsections</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial paragraph</strong></td>
<td>State the main findings of the study, which were raised as objectives in the Introduction section.</td>
</tr>
<tr>
<td><strong>Body paragraphs</strong></td>
<td>Establish the novelty or uniqueness of the results of the paper.</td>
</tr>
<tr>
<td></td>
<td>Support your study results by comparison with results published by other groups of investigators. Cite other studies’ results that support or are discordant in relation to your study results explaining how they fit or conflict.</td>
</tr>
<tr>
<td></td>
<td>Defend your results generating hypotheses for why the answers of your study are different from the answers from previous published studies.</td>
</tr>
<tr>
<td></td>
<td>Discuss unexpected results that do not support the hypotheses of your study.</td>
</tr>
<tr>
<td></td>
<td>Discuss the strengths, weaknesses, and limitations of your study, indicating how design limitations of your study may have led to over- or underestimation of results’ values.</td>
</tr>
<tr>
<td><strong>Final paragraph</strong></td>
<td>Restate the answer to the primary research question and indicate the significance of the study results with regard to clinical applications, implications for the health care system, and recommendations.</td>
</tr>
<tr>
<td></td>
<td>Suggest future needs for investigation and unanswered questions.</td>
</tr>
<tr>
<td></td>
<td>End with a strong, clear, evidence-supported conclusion (“Summary Statement”).</td>
</tr>
</tbody>
</table>

### Introduction Section

Only strictly pertinent background information that informs the reader as to why the study was performed should be included in the Introduction section. Brevity and focus are key. No detailed review of the topic is expected in this section. What is said in the Introduction section does not need to be repeated in the Discussion section. An effective Introduction has three segments: a section that provides a scoping statement (why the authors decided to investigate the proposed research question[s]), a section that focuses somewhat on the research area (what is currently available in terms of technology, processes, and procedures to address the question[s]), and a section that focuses on the specific area of research for the current study (how the proposed method or strategy can potentially be superior to the methods or strategies currently available).24 In the final paragraph of this section the authors should state the objectives/hypotheses of the study. For example, for an Original Research manuscript that allows a maximum of 3,000 words, the information in this section is expected to not exceed 400 words.21 Details of the Introduction section are available in Table 10.4.
Conducting and Publishing Research

10

If the authors have used a reference management software program during the writing of the manuscript, then the software will number the references as they are inserted in the manuscript and format the references according to the style requirements of a particular journal or publisher. Otherwise, authors should manually number references consecutively in the order in which they are first mentioned in the manuscript for most journals, although a minority of journals may use alphabetical order. Articles that appear in online journals should follow the same citation format as print articles, with the addition of the URL and the date the article was accessed. In the case of books, the authors of a chapter, title of the chapter, editor(s), title of the book, edition, city and state, publisher, year, and specific pages must be provided. For web content, author(s) (if any); title of the page or content; name or owner of the website; URL; and publication, update, and access dates should be listed. The authors should search for types of references in the online Publication Information for Authors of the proposed journal. Most journals state that it is the responsibility of the authors to verify the accuracy of their study references to ensure proper linking of referenced articles in the online journal.

Abstract Section

Once the four sections of the manuscript (Materials/Patients and Methods, Results, Discussion, and Introduction) are completed, the author should re-read them and draft the Abstract section. Therefore, it is written last, not first. For Reviews, Original Articles, or other similar submissions, either structured or unstructured abstracts (typically 100–300 words depending on the journal) should summarize the content of the submission, and specific headings (Objectives, Methods, Results, and Conclusion) should or should not be included as per guidelines of the scientific journal (Table 10.5). Authors should include key words (typically an average of six; no less than four and no more than eight is a good guideline) for the study on a subsection of the abstract. Key words should be concise and authors should avoid abbreviations. For radiology studies, at a minimum, investigators should include keywords identifying the patient age group, the imaging modality studied, the organ system, and the disease process.

References

Reference formatting is not the same for all journals. In addition to reading instructions, authors may look at articles from the same journal as examples for proper formatting. If the authors have used a reference management software program during the writing of the manuscript, then the software will number the references as they are inserted in the manuscript and format the references according to the style requirements of a particular journal or publisher. Otherwise, authors should manually number references consecutively in the order in which they are first mentioned in the manuscript for most journals, although a minority of journals may use alphabetical order. Articles that appear in online journals should follow the same citation format as print articles, with the addition of the URL and the date the article was accessed. In the case of books, the authors of a chapter, title of the chapter, editor(s), title of the book, edition, city and state, publisher, year, and specific pages must be provided. For web content, author(s) (if any); title of the page or content; name or owner of the website; URL; and publication, update, and access dates should be listed. The authors should search for types of references in the online Publication Information for Authors of the proposed journal. Most journals state that it is the responsibility of the authors to verify the accuracy of their study references to ensure proper linking of referenced articles in the online journal.

Table 10.4 Suggestions for drafting the Introduction section of a manuscript

<table>
<thead>
<tr>
<th>Introduction subsections</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial paragraph</td>
<td>Define the disease or problem the authors decided to evaluate and its implications on the morbidity and mortality for patients.</td>
</tr>
<tr>
<td>Body paragraphs</td>
<td>Address the problem or gap in the literature and state why the authors have decided to conduct this research study and why it is so important to fill in these gaps.</td>
</tr>
<tr>
<td></td>
<td>Comment on the diagnostic approaches that are currently available to fill in the literature gaps, how did the authors plan to do so in the study, and what imaging technology they have used in their research study. If the technology is novel, they should briefly explain what the technique is about, stating advantages over conventional techniques.</td>
</tr>
<tr>
<td></td>
<td>Link the problem with the technique that is available to diagnose the problem (disease). In this paragraph, the authors can briefly report results of other studies in the field that provided incomplete information on the technique–problem linkage and how the current study will be complementary to the questions raised. Authors should also state whether no previous study has addressed the proposed research questions of their study, which is indeed needed for this study to be relevant.</td>
</tr>
<tr>
<td>Final paragraph</td>
<td>Clearly state the hypothesis and purpose of the study in a fashion similar to the Purpose statement in the abstract.</td>
</tr>
</tbody>
</table>
Table 10.5  Examples of radiology journals’ website instructions for authors concerning original research manuscripts’ abstract format, total number of words for abstract and full manuscript, and maximum number of figures, tables, references, and authors

<table>
<thead>
<tr>
<th>Journal</th>
<th>Type of abstract and specific requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Brain Mapping</td>
<td>Structured: objectives, experimental design, principal observations, and conclusions</td>
</tr>
<tr>
<td>Radiology</td>
<td>Structured: purpose, materials and methods (compliance [IRB, HIPAA, informed consent, animal use committee]), results, conclusion</td>
</tr>
<tr>
<td>Neuroimage</td>
<td>Nonstructured synopsis: purpose, principal results, and conclusions</td>
</tr>
<tr>
<td>Journal of Nuclear Medicine</td>
<td>Structured: rationale, methods, results, and conclusions</td>
</tr>
<tr>
<td>Investigative Radiology</td>
<td>Structured: objectives, materials and methods, results, and conclusions</td>
</tr>
<tr>
<td>European Journal of Radiology</td>
<td>Nonstructured synopsis</td>
</tr>
<tr>
<td>Magnetic Resonance in Medicine</td>
<td>Structured: purpose, methods, results, and conclusion</td>
</tr>
<tr>
<td>American Journal of Neuroradiology</td>
<td>Structured: background and purpose, material and methods, results, and conclusion</td>
</tr>
<tr>
<td>American Journal of Roentgenology</td>
<td>Structured: objective, materials (or subjects) and methods, results, and conclusions</td>
</tr>
<tr>
<td>Pediatric Radiology</td>
<td>Structured: background, objectives, material and methods, results, and conclusion</td>
</tr>
<tr>
<td>Academic Radiology</td>
<td>Structured: rationale and objectives, materials and methods, results, and conclusion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Journal</th>
<th>Abstract (words)</th>
<th>Full manuscript (words)</th>
<th>Figures and tables</th>
<th>References</th>
<th>Number of authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Brain Mapping</td>
<td>250</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>Radiology</td>
<td>250</td>
<td>3,000</td>
<td>8 figure parts, 4 tables</td>
<td>35</td>
<td>Not specified</td>
</tr>
<tr>
<td>Neuroimage</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>Journal of Nuclear Medicine</td>
<td>350</td>
<td>5,000</td>
<td>7 figures, 7 tables</td>
<td>40</td>
<td>Not specified</td>
</tr>
<tr>
<td>Investigative Radiology</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>European Journal of Radiology</td>
<td>200–250</td>
<td>4,000</td>
<td>Not specified</td>
<td>30</td>
<td>Not specified</td>
</tr>
<tr>
<td>Magnetic Resonance in Medicine</td>
<td>200</td>
<td>5,000</td>
<td>10 (figures plus tables)</td>
<td>Not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>American Journal of Neuroradiology</td>
<td>250</td>
<td>4,500</td>
<td>Space taken up by figures is accounted for according to the journal’s image and word count guidelines</td>
<td>Not specified</td>
<td>5–10 for retrospective studies and 10–15 for prospective studies</td>
</tr>
<tr>
<td>American Journal of Roentgenology</td>
<td>250</td>
<td>4,500</td>
<td>7 (15 figure parts), 4 tables</td>
<td>50</td>
<td>7</td>
</tr>
<tr>
<td>Pediatric Radiology</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>Academic Radiology</td>
<td>250</td>
<td>4,000</td>
<td>4 figures, 4 tables</td>
<td>20</td>
<td>Not specified</td>
</tr>
</tbody>
</table>

Abbreviations: HIPAA, Health Insurance Portability and Accountability Act; IRB, institutional review board.
Tables

Tables and figures can be prepared during drafting of other sections or at the end of the writing process. This process requires accuracy in that information of tables matches with that reported in the text of Results. Tables should be numbered in Arabic numerals and should have a title. Each table should have sufficient information that allows it to stand by itself. All abbreviations used in the table should be explained in a footnote. Abbreviations already explained in the manuscript text should be repeated in the labels or footnotes of tables and legends of figures. The authors should provide the unit of measurement for each variable, report counts along with percentages, specify the type of measure of central tendency (mean, median, etc.), and measure of dispersion (e.g., standard deviation, standard error, interquartile range). For results of comparisons, authors should specify the summary measure (mean difference, relative risk, odds ratio, etc.) and provide a measure of precision around the estimate (e.g., 95% confidence intervals). Some journals accept supplemental material, which can be uploaded as an Appendix on the journal’s website.

Illustrations

Figures provide visual aids in the form of drawings, surgical photographs, and imaging captions to facilitate the reader understanding on the research or teaching point. They should appear sequentially in the text supporting information reported in Results, and less likely, in Discussion or Materials/Patients and Methods. A legend must be supplied for each illustration, including drawings and graphs, and should not duplicate text or table material. It is essential that the legend describes all labels and annotations placed on an illustration. Judicious use of annotations is recommended, enough to assure reader understanding, but not too many as to be distracting. Use of simple annotations (black and/or white, not colored) is suggested. Legends describing radiologic images should include the type of image, its imaging projection or plane, use of contrast material, pulse sequence information (repetition, echo, and inversion times; flip angle) for magnetic resonance images, as appropriate, and the features to be observed by the reader highlighted by arrows and other indicators (Fig. 10.1). Preferably, patient age/gender is indicated in the legend. Regarding graphs, the x-axis and y-axis are labelled including units, and all components

Fig. 10.1 Examples of visual aids in the form of drawings, surgical photographs, and imaging captions to facilitate the reader’s understanding of the research or teaching point. Fourteen-year-old teenager with surgically diagnosed uterus didelphys with obstructed right hemi-vagina and ipsilateral renal agenesis. (a) Drawing of this patient’s Müllerian duct anomaly type. Red shading represents the internal aspect of each uterus didelphys part. Arrow shows the obstructed right hemi-vagina. (b) Axial T2-weighted fast spin-echo (FSE) MRI. There is a large hematocolpos (arrow) centrally placed, however, corresponding to the right-sided obstructed hemi-vagina. (c) MRI, coronal plane. Solitary left kidney. (d) Surgical picture shows a bulging right vagina (arrow). (Reproduced with permission from Junqueira BL, Allen LM, Spitzer RF, Lucco KL, Babyn PS, Doria AS. Müllerian duct anomalies and mimics in children and adolescents: correlative intraoperative assessment with clinical imaging. Radiographics 2009;29:1085–1103.)
Research Methods in Radiology

should have a clear vision of their entire work before promoting it. The main purpose of the letter should be to convey any information that the editor might need to know. Usually reviewers do not see the cover letter. This letter could be written on a single page to the editor of the journal to which the manuscript will be submitted. An example of a cover letter is available in Fig. 10.3. The authors should briefly tell the journal’s editor why it is so important for the journal to accept their manuscript. In the cover letter authors should also state that the submission represents original work not previously published in any substantial part, if this information is true; that it is not under consideration of publication elsewhere; and that it has been submitted to the journal for consideration of publication. If any portion of the patient population has been used in another study, this needs to be explicitly stated and explained (with references, if applicable). It is suggested that the authors of the graph should be adequately explained in the legend. If flow diagrams of research designs or protocols are used, they should indicate the events in a chronologic fashion, with the earlier events being displayed on the left-hand side of the figure and the later events on the right-hand side (Fig. 10.2). Concerning photomicrographs and pathology slides, the authors should include the stain and original magnification. Figures should be provided in high resolution according to the journal’s instructions. Most journals require digital images of at least 300 dpi (dots per inch; 1200 dpi for line art). Specifications regarding image preparation can be found in the online Publication Information for Authors of journals. Color figures may require the payment of a fee from the authors for paper copy publication.

Cover Letter to the Editor

This should be the last step of the manuscript writing process because the authors of the graph should be adequately explained in the legend. If flow diagrams of research designs or protocols are used, they should indicate the events in a chronologic fashion, with the earlier events being displayed on the left-hand side of the figure and the later events on the right-hand side (Fig. 10.2). Concerning photomicrographs and pathology slides, the authors should include the stain and original magnification. Figures should be provided in high resolution according to the journal’s instructions. Most journals require digital images of at least 300 dpi (dots per inch; 1200 dpi for line art). Specifications regarding image preparation can be found in the online Publication Information for Authors of journals. Color figures may require the payment of a fee from the authors for paper copy publication.

Fig. 10.2 Flow diagram of the experimental protocol of a hypothetical study. Patients with clinical suspicion of a tumor have undergone a conventional MRI for clarification as per clinical practice. If the diagnosis of a new sarcoma was confirmed by biopsy and patients met other inclusion criteria for the study, the patients/families were consulted about participation in the study. Consenting patients underwent blood oxygenation level–dependent (BOLD) MRI and MR spectroscopy (MRS) within a week from the biopsy results. Patients received treatment for their cancer as per clinical practice. Prior to the resection of the tumor, participants underwent follow-up BOLD MRI and MRS. A time frame no longer than 4 weeks was expected between follow-up imaging and surgery for tumor resection. The resected tumor specimen was sent off to pathology for further analysis.
Re: Manuscript “Diagnostic Accuracy of Ultrasound for Assessment of Hemophilic Arthropathy. MRI Correlation”

Please find enclosed the manuscript entitled “Diagnostic Accuracy of Ultrasound for Assessment of Hemophilic Arthropathy. MRI Correlation” for consideration by your Journal.

In this manuscript we provide detailed soft tissue and osteochondral ultrasound (US) and MRI diagnostic test information on pathologic ankles and knees of 59 boys (< 18 years of age) with hemophilia / von Willebrand Disease enrolled in a cross-sectional study conducted in a single center.

Our findings are novel and relevant for several reasons. Key points are as follows:

With special regard to developed countries where diagnostic imaging focuses on the assessment of “minimal changes” and musculoskeletal clinical trials aim to use imaging outcomes to measure “minimal interval changes” in response to therapy, the results of both manuscripts are paramount. As an example the definition of failure in the randomized, controlled joint Outcome Study reported by Manco-Johnson and colleagues in the New England Journal of Medicine (Manco-Johnson MJ et al. N Engl J Med 2007;357:535-544) comparing primary prophylaxis to episodic (“on-demand”) therapy was based on presence of osteochondral abnormalities detected by MRI. Without a clear definition of normal US and MRI findings on growing joints of children and adolescents of different age groups, it is not possible to discriminate minimal arthropathic changes from normal physiologic growth patterns of joints. This could lead to errors while evaluating minimal and mild joint changes by US and MRI in the pediatric population.

The results of the "pathologic group" study rely on the “best case scenario” (unblinded readers and US operators) to evaluate the diagnostic accuracy of US as compared to MRI for the evaluation of soft tissue and osteochondral changes. As a result, the study design contains expected biases related to unblinding of readers and US operators. Nevertheless, the study represents the “back bone” of scientific knowledge to enable future advancement of the field. Only after atlases of normative and pathologic US and MRI findings for children of different age groups become available in the literature will investigators be able to fully determine the “real life” blinded diagnostic accuracy of US for assessment of musculoskeletal disease in maturing joints.

The results of the two studies have clinical applications not only for hemophilic arthropathy, but also for all other chronic arthropathies in the pediatric population.

The following information is available as Supplemental material:

- Table 1, Supplemental material: International Prophylaxis Study Group MRI scale for interpretation of findings in examinations of ankles or knees.
- Table 2, Supplemental material: data representation of US and MRI scores in examinations of ankles (a) and knees (b) according to the items of each scale.
- Table 3, Supplemental material: Frequency of findings and true-positive, true-negative, false-positive, false-negative values for ankles and knees.

For the aforementioned reasons we believe that the results of our prospective studies deserve publication and have selected your journal for consideration of our work because of its international stature and global readership.

[If requested by the journal] We suggest the following individuals as potential reviewers for our work:

1. Dr. Y
2. Dr. Z
3. Dr. W

Many thanks for your consideration of our work.

Yours sincerely,

Andrea S. Doria, MD, PhD, MSc (Corresponding Author)
Professor, Associate Vice-Chair of Research (Inflammation, Repair and Injury),
Department of Medical Imaging, University of Toronto
Research Director, Department of Diagnostic Imaging
Clinician-Scientist, Radiologist
The Hospital for Sick Children
555 University Avenue, 2nd floor
Toronto, ON M5G1X8
Phone: 416-813-6079
Fax: 416-813-7591
Email: andrea.doria@sickkids.ca

Fig. 10.3 Example of a cover letter for submission of a manuscript for consideration of publication in a scientific journal.
indicate if the study has been presented at a scientific meeting prior to consideration for publication.

**Title Page**

This page should include title, authors’ names (including name and address of the corresponding author and of the institution from which the work was originated), and funding information. Most journals require that at least one author acts as a guarantor of the work, and takes responsibility for the publication as a whole.

The title of the manuscript should be simple and explicitly summarize what the authors did in as few words as possible. The reader should have a basic idea about what was done in the project from just the title. Not uncommonly this is all that will be read about the paper in PubMed.

At the time of finalization of manuscript drafting (preparation of the title page), the PI should revisit the a priori defined list of authors to ensure that all the authors have successfully accomplished the tasks assigned to them in the beginning of the study, thus securing their expected authorship position in the manuscript.

**Acknowledgments**

As stated by Bourbonniere et al., contributionship aims to diminish ambiguity and encourage transparency and responsibility as the specific contributions of all individuals named in the byline are disclosed to the reader.

The Acknowledgment subsection can include names of individuals who collaborated with the conduct of the study but did not meet authorship/contributorship criteria, information about prior or future (upon proof of acceptance) presentation of the work in scientific meetings, and source of funding obtained. Concerning acknowledged contributors, because readers may infer their endorsement of the data and conclusions, these persons must give written permission to be acknowledged.

**Additional Submission Requirements**

The authors should state the type of manuscript being submitted. Keywords can be taken from the Medical Subject Headings List. Some journals may require descriptions on word count for the Abstract and Main Manuscript Body and justification of individual author contributions for the study. Whereas some journals require the signatures of all authors for Copyright Agreement by the time of submission of manuscripts, others only required this after the manuscript has been accepted for publication. This information should be available on the Instructions for Authors website of the journal. Certain journals (e.g., Radiology) may have special components such as the request for a “Summary Statement,” which is a single sentence taken directly from the text that best summarizes the manuscript.

Most journals request suggestions for reviewers (variable number according to the journal) from the authors at the time of the manuscript submission. The journal trusts the authors’ judgement as to who the experts in the area of research interest of their study are. Preference should be made to professionals with different backgrounds related to the topics of investigation reported in the manuscript.

**Open Access Publications**

Eysenbach showed that open access articles compared to non-open access articles remained two to three times as likely to be cited 4–16 months after publication. In addition, open access articles are more immediately recognized and cited by peers than non-open access articles published in the same journal. Thus, open access to the research literature has the potential to accelerate recognition and dissemination of research findings, but its actual effects are controversial. This approach is highly encouraged by grant agencies to increase the Knowledge Translation effect of the study results. With Online Open, the author, the author’s funding agency, or the author’s
institutions pay a fee to ensure that the article is made available to nonsubscribers. Note should be made that in some circumstances the authors are under obligation to publish open access (i.e., National Institutes of Health or Canadian Institutes of Health Research funded, work in certain countries, etc.).

**Submission of the Manuscript for Consideration of Publication**

**Presentation of Results at Scientific Meetings**

Many investigators make the common (and unfortunate) mistake of leaving aside their manuscript after having completed a substantial amount of work. It is surprising to learn that only approximately 63% of authors tend to submit manuscripts for publication after their presentation at scientific meetings. The average time to publish an article after its presentation is 20 months. Presentation of results at a scientific meeting is part of the process but not the final goal of publishing an article. Ideally authors should write up their manuscript in conjunction with assembling the corresponding scientific presentation. In doing so, the authors can have a better understanding of their work, particularly the limitations. This will make for a better scientific presentation and equip the authors to better answer questions from the audience at the time of presentation at the scientific meeting. Alternatively, after returning from the scientific meeting where the work was presented, the authors should focus on submission of their manuscript to a journal. Commonly authors are reluctant to attempt to publish negative results, where an expected outcome did not happen, as they believe that these negative findings are not as important as positive ones. This is a mistake that may result in publication bias. This bias occurs when the publication of research results depends on their nature and direction and may cause a shift in the results of meta-analyses or systematic reviews, because current data tend to overestimate diagnostic accuracies and treatment effects when not all authors publish their results. Both positive and negative results of studies conducted with high methodologic standards should receive encouragement for submission for consideration of publication.

**Manuscript Review by Authors**

The first and/or senior author should make modifications in the manuscript as per suggestions of peers in the scientific meeting (if presented), proofread the entire document, and circulate this draft among the authors for internal review. Once the leading author receives the comments and suggestions of the other authors, he/she should make changes to the manuscript and recirculate it to all contributing authors for final review.

**Electronic Submission**

Although authors may start discussing potential scientific journals to which their study results can be submitted by the time the first manuscript draft is prepared the final decision on the journal for submission is usually made once the first draft of the manuscript is completed. This strategy enables authors to reevaluate the overall quality of their final product, which should direct the journal to which material should be submitted based on the expected impact of the study results and readership of interest. In addition, other factors may play a role such as length of the manuscript (some journals, e.g., *New England Journal of Medicine* and *Radiology*, have strict rules on maximum number of words for original research contributions, 2,500 and 3,000, respectively), maximum number of authors, turnaround time for peer-review process, and publication (which is critical for ground-breaking results), among others.

Finally, the authors should upload the manuscript documents onto the journal’s website and submit the materials. The corresponding author should receive an electronic notification from the journal assigning a manuscript identification/number, which should be used for future reference.
Peer Review and Results of Consideration for Publication

The peer review process ensures that each article conveys its message accurately, unambiguously, based on convincing results. In a revision, the authors should carefully address all comments and suggestions from the reviewers and editors. A cover letter should explain the revisions. If a change is not made in response to any comment or suggestion, the reason for not making a change should be explicitly explained.

The turnaround times of scientific journals for peer review may vary, from weeks to months depending on the journal. Once a decision on the status of the manuscript is reached, the authors should receive an email notification from the journal’s editor indicating whether the manuscript was rejected, accepted with no revisions (which is rare), accepted with minor revisions, or accepted with major revisions upon satisfactory fulfillment of points raised by the reviewers. If the manuscript was rejected the editor usually indicates that the work cannot be resubmitted to that journal. Reasons for a manuscript rejection include poor study design, lack of new knowledge, inadequate description of methods, suboptimal reporting of results, poor discussion, poor writing, lack of compliance to the journal’s guidelines for submission, and submission of the manuscript to an inappropriate journal (topic of research is beyond the journal’s scope of interest). Also, failure to revise and resubmit after peer review constitutes a reason for manuscript rejection. It is acknowledged that replying to the editor’s and reviewers’ solicitation of resubmission of the manuscript in case of need for major or minor revisions improves the likelihood of publication of the manuscript.

Many investigators follow the strategy of first submitting their manuscript to the journal in which they would ideally like to see their work published, and if rejected, to the journal that will most likely accept their work. If the authors choose to resubmit their manuscript to another journal, they should (1) revise it in accordance to the comments/suggestions obtained from the journal that rejected the submission; and (2) reformat their manuscript in accordance with the author instructions of the new journal.

Review of Galley Proofs of Accepted Publications

Upon variable time after the manuscript acceptance for publication, the journal editorial office emails the corresponding author electronic page proofs (galley proofs), usually as a PDF that resembles the final journal pages. Queries received from the publisher or as part of the galley proofs should be electronically answered in a timely fashion. It is the responsibility of the authors that the proofs are accurate and error-free. The allowable changes are mainly typographical.

Frequently Asked Questions

Answers to these frequently asked questions are available in the supplementary material.

Authorship

What should be the order of authors in a manuscript?

Is conjoint first authorship possible?

Knowledge Translation

Should investigators present their work in scientific meetings prior to its publication?

What is the most suitable scientific meeting for presentation of the investigators’ work?

Presentation of Abstracts in Scientific Meetings

Should the Abstract submitted for presentation in a scientific meeting comprise the same sample size of the manuscript submitted for publication?

Preparation of Manuscripts

Can investigators split their study results into two or more manuscripts?
## Highlights of Key Points

- There is a basic recipe to draft a manuscript that should be adjusted to the availability of time and level of experience of authors. Nevertheless, grouping academic days consecutively, especially for drafting the Discussion section, may facilitate recalling information from the literature for comparison of findings of the study with relevant results of other studies. The Materials/Patients and Methods section can be outlined during the initial phase of the study and refined upon its completion.

- It is highly recommended that authors follow the instructions of the journal selected for submission of the manuscript, and that review guidelines from Standards for Reporting of Diagnostic Accuracy (STARD), Consolidated Standards of Reporting Trials (CONSORT), and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) be used as checklists for excellence on reporting of studies dealing with diagnostic accuracy, randomized controlled trials, and meta-analyses of randomized controlled trials, respectively.

- Results section of a manuscript should mirror the Materials/Patients and Methods section, including same order, and should not include anything that was not already introduced in the Materials/Patients and Methods section. No results should first appear in the Discussion section.

- Consultation with a statistician before the commencement of a study is primordial to ensure that its sample size will provide enough power, to discuss the database where the data will be entered (this strategy will save time in the future by preventing the need for major modifications), and to confirm that the methods used for analysis will be able to answer the proposed research questions.

- A power literature search may require the guidance of an experienced librarian. It should be performed by the time the investigators are determining gaps in the literature and the value of the proposed research question and updated by the time the authors plan to draft the Discussion section of the manuscript.

- It is recommended that the authors create a reference library using an appropriate software at the time of the initial literature search to facilitate management of references later on during the manuscript draft phase.

- Publications are a measure of a researcher's productivity and are used when making decisions on applications for new jobs, academic promotion, or tenure track. For this and other reasons, authorship is an important issue on publications that warrants discussion and agreement on clear guidelines early on during the conduct of the study and should be a regular agenda item at research project meetings. This should avoid future conflicts around authorship issues. Different schemes for weighting authors' contributions are described in this chapter.

## Tips to Improve the Likelihood of Publication of a Manuscript in a Scientific Journal

1. Organization of the manuscript parts, and coherence of ideas throughout the text.
2. Clarity of description of research question(s), study rationale, methods, report, and discussion of results.
3. Systematic structure of results and discussion sections: correspondence of points under investigation between these two sections.
4. Explanation about why the study results are important and how they can be clinically applicable.
5. Inclusion of one paragraph in the discussion section about study limitations (design, sample size, methodology) and how the authors have tried to overcome them.
6. Description of unexpected results and how they have been handled by the authors.
7. Good presentation of tables and figures, including consistency of proper significant digits for numbers, percentages, etc., and for alignment of rows and columns in tables.

8. Compliance to the guidelines of the journal as related to maximum number of words, tables, figures, and authors, if appropriate.

9. Reply to the editor's request for "major" or "minor" revisions within the time frame provided by the journal in the email response from the editor if the authors are offered the opportunity of resubmitting the manuscript to the journal to which they had originally submitted the manuscript.

10. Acceptance of reviewers' comments and suggestions if appropriate. If not, the authors should draft a strong justification about why they have decided not to follow the reviewers' suggestions.

References


15. Flanagin A, Fontanarosa PB, DeAngelis CD. Authorship for research groups. JAMA 2002;288:3166–3168


27. Eysenbach G. Citation advantage of open access articles. PLoS Biol 2006;4:e157
30. Krzyzanowska MK, Pintilie M, Tannock IF. Factors associated with failure to publish large randomized trials presented at an oncology meeting. JAMA 2003;290:495–501
31. Pierson DJ. The top 10 reasons why manuscripts are not accepted for publication. Respir Care 2004;49:1246–1252
32. Provenzale JM. Ten principles to improve the likelihood of publication of a scientific manuscript. AJR Am J Roentgenol 2007;188:1179–1182
## 11 Obtaining Ethics Approval, Data Management, and Budget Formulation

*Ravi Menezes, Yasser Karimzad, and Tuula Kalliomäki*

### Learning Objective

The objective of this chapter is to introduce interested but inexperienced researchers to a number of practical concepts that are important to the planning, design, and conduct of a research study. On the surface, institutional approval for conducting a study, data management, and budget formulation may not appear to be directly linked. Closer examination, however, reveals that they are strong determinants of the day-to-day operations of both large and small-scale research studies. These important pieces of a study are intended to ensure that it adheres to institute guidelines (application for institutional research approval), the data is accurate and credible (data management), and the appropriate resources are available to conduct the study as planned (budget formulation). Each of these topics is covered in this chapter, largely presented conceptually with practical advice for what to do or avoid. Where possible, issues of particular relevance to radiology research are highlighted.

As is mentioned throughout the chapter, readers are encouraged to investigate the regulations and restrictions that affect the research activities at their local institutions and from relevant sponsors and granting agencies. A summary of key points to remember for each of the three sections is included at the end of this chapter.

### Obtaining Ethics Approval and Informed Consent

The objective of clinical research is to advance the knowledge and understanding of human health and biology. Individuals who consent to be research participants usually do not benefit directly from the research despite the potential for risk. Guidelines for the ethical conduct of clinical research exist to minimize the chance of exploitation of research participants by assuring that their rights are respected. This section provides an introduction to the role these guidelines play in clinical research and highlights important aspects of the independent ethics review and informed consent. Since the specific details of these processes vary by institution, it is recommended that readers familiarize themselves with the rules that govern the approval and conduct of research at their local institutions.

### Important Ethics Guidelines

Since the Second World War, numerous guidelines have been produced to address ethical issues in human research participants. The purpose of the majority of these guidelines is to ensure the protection of the rights of research participants and to guarantee data integrity. Detailed review of these important guidelines is beyond the scope of this chapter, however, and interested readers are encouraged to look up details related to the following: the Nuremberg Code, Declaration of Helsinki, the Belmont Report, United States’ Code of Federal Regulation (CFR), regulated by the U.S. Food and Drug Administration, the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), and Guideline for Good Clinical Practice (GCP). These guidelines have played an integral role in the creation of rules that govern research conduct at institutions throughout the world, and should be followed when generating clinical trial data that are intended to be submitted.
to regulatory authorities. The principles established in these guidelines may also be applied to other clinical investigations that have an impact on the safety and well-being of human participants. It is strongly recommended that researchers become familiar with the rules and regulations that govern research at their local institutions.

**Institutional Review Board**

GCP guidelines require that all research involving human participants be reviewed and approved by an Institutional Review Board (IRB), also known as the Research Ethics Board (REB) or Independent Ethics Committee (IEC). The IRB consists of scientific and nonscientific members responsible for ensuring the protection of the rights and safety of human research participants. Regulatory authorities have given IRBs the power to approve studies, request for changes in planned research before approval, or reject research proposals. The IRB is also responsible for paying extra attention to vulnerable research participants, such as pregnant women, children, the elderly, mentally disabled persons, and members of groups with a hierarchical structure, such as medical and nursing trainees. It is imperative that the appropriate IRB has reviewed and approved a research study before any intervention or interaction with human research participants.

**Required Documentation**

The IRB will require specific documents to be submitted for review. These documents will vary by study design and the specific IRB from which approval is being requested. This is an important reason why investigators must familiarize themselves with their IRB guidelines prior to the submission of documents for approval.

A majority of IRBs will request a specific application form to be completed and a research protocol to serve as the backbone of the application. The protocol should address the project summary, study rational, background information, study design, methodology, hypothesis, statistical details, and participant safety considerations. The ICH guidelines provide excellent recommendations regarding research protocol contents.

For prospective studies that require the recruitment of patients, an Informed Consent Form (ICF) is essential. The basic elements of an ICF are generally standard and similar in all the ICH regions around the world and are described in this chapter. Many IRBs will have a suggested template available for modification by researchers. Most clinical trials also require the submission of other documents such as an Investigator’s Brochure (IB), study budget, data collection forms, and participant recruitment procedures and tools, such as advertisements and telephone follow-up scripts. Depending on the complexity of the study and where it is to be conducted, additional approvals, certifications, and contracts may have to be obtained from different hospital departments, governmental authorities, and standardization associations.

Industry-sponsored studies usually require full contracts executed by the institution’s contracts and legal department before approval is granted. For multicenter studies initiated by a principal investigator (PI), the IRB may request a data sharing agreement to be in place between the lead site and other participating centers. The main purpose of this agreement is to assure that patient confidentiality rules and regulations are followed and that all data and documents shared between sites are fully anonymized.

**Informed Consent**

Informed consent has been a cornerstone of ethically conducted research dating back to the Nuremberg Code, the first formal research guideline. Prior to any research-related procedure, consent must be obtained and an ICF must be signed by the participant and the consenter. It is vital to remember that the form is not a contract. By signing the form, the participant does not give up any of his/her rights and is able to withdraw from the study at any time. In certain study designs, especially those addressing inter- and intrareader agreement, it might be necessary to identify radiologists as research participants. In these
Institutional Research Board Review of Retrospective vs. Prospective Studies

Both retrospective and prospective study designs require IRB approval prior to initiating the study. In retrospective studies, investigators attempt to answer study questions using data that have already been collected by the time the IRB application has been submitted. The data, which can come from the radiology Picture Archiving and Communication System (PACS) or patient medical records, may have been generated for reasons other than research. Depending on the design and objectives, retrospective studies may not require full board approval and their review will be expedited by delegates of the IRB chair. Prospective studies are designed before any data is collected. Specific study participant inclusion and exclusion criteria are set and participants are followed during the course of a defined study period. A majority of prospective studies will go through a full board approval where all members of the IRB must review the study. The scope of the review is determined by the IRB and not the researcher.

Data Confidentiality

Personal health information (PHI) includes data that can be used to identify participants on their own or in combination. Examples include, but are not limited to, names, medical record numbers, and accession numbers. IRBs often place restrictions on the use of PHI, and it is the researchers’ responsibility to keep this data confidential if it is collected. The investigator must assure the IRB that information related to the participant will be respected and not divulged to others without permission, providing details that describe the responsible collection and storage of data. Such assurances should also be described in the ICF in more general terms, as this helps establish trust between researchers and participants.

As good practice, it is often recommended by IRBs that each enrolled participant be assigned a unique study identifier (study ID). If documents are updated and a new version of the consent becomes available, the researchers must ensure the most recent version of the form is used.
programs. This term includes a definition for the file format and a protocol for network communications. For radiology studies where DICOM images are to be shared between sites, it is recommended that all images be fully anonymized using pretested and approved methods. This may be mandatory at some institutions.

**Data Management**

Data management refers to study activities related to the collection and storage of data. Because of the direct link to study design, it is important that study objectives and design details be documented prior to planning data-related activities. This will increase the likelihood that consensus and clarity will drive subsequent study-related decisions. It is also important that data management activities, especially the flow of information and the personnel and equipment required to carry them out, be determined as early as possible. These will likely need to be authorized by an IRB prior to the commencement of the study.

The importance of good data management practices cannot be overstated. High study integrity and data transparency are important to maintain. More practical reasons for good management practices include easy communication between study team members, less time redoing tasks or correcting errors, being able to hit submission targets for conferences or grants, and more efficient analysis. The contents of this section describe concepts and make suggestions that may help principal investigators and the study team achieve the intended study goals.

**Defining Data Elements**

On the collection end, data elements should be defined. These elements include units of measure, precision, and scales. It is recommended that commonly accepted terminology and measurements be weighted heavily as they improve generalizability of the data and the eventual conclusions that will be made. Useful sources for such terminology are relevant clinical consortiums and organizations. Since the publication of results and integration of findings into clinical practice are two frequent and important goals of research, the use of common terms and data elements will provide a better context to outside reviewers and readers as to what the study has accomplished and how it can be translated in the local or external institutions.

Most study designs should involve data that are objective and do not require interpretation. This means that the variables can be clearly defined to study team members and an external audience and are easy to understand and replicate during the conduct of the study as well as in follow-up studies.

**Data Dictionary**

A data dictionary, also known as a glossary, is a document that lists and describes all data elements. It is an important tool for a number of reasons; most obvious is keeping track of what data are being collected. It should include variable descriptors such as format (e.g., string, numeric, categorical, date, etc.), category definition, coding scheme, and the definition of the variables. The dictionary can be maintained in list or table format in a word processor or spreadsheet program, with each descriptor listed as a field or column heading. The document is especially useful if a statistician is responsible for the data analysis.

**Data Collection**

The data collection process is of utmost importance to study integrity. While data sources will differ from primary (imaging studies, study participants) to secondary (patient records), similar data collection methods and instruments can be applied for all types of studies. The process and data should be unbiased and objective, have a low likelihood of error, and result in data that are easy to analyze.

**Data Collection Forms**

Data collection forms (DCF), also known as case report forms (CRFs), are commonly used programs. This term includes a definition for the file format and a protocol for network communications. For radiology studies where DICOM images are to be shared between sites, it is recommended that all images be fully anonymized using pretested and approved methods. This may be mandatory at some institutions.
tools in clinical research. They can be paper or electronic, and consist of a collection of clearly identified fields that guide the user to search for and record relevant data. DCFs are especially useful if multiple people are involved in the data collection or generation stage, including study coordinators, research assistants, and radiologists. The template format prompts readers to record data in a consistent fashion, and the series of completed forms provides an audit trail.

DCFs should be designed so that they are self-explanatory and require minimal interpretation by users. Units of measure and levels of precision should be indicated on the forms, especially for fields in which this can be variable (e.g., dates, anthropometrics). The inclusion of features such as checkboxes and options items to circle will reduce the need for handwritten responses and likelihood of error. The sequence and relative spacing of items in the horizontal and vertical axes of the DCFs should be chosen to minimize the moving of the eyes across a form as the user searches for and fills in a field. This can be accomplished by matching the layout to the natural order with which observations are recorded when radiologists interpret images or review reports.

**Test Runs**

Data collection and entry should be tested prior to full study rollout. The same individuals who will perform these activities for the actual study should be involved in the testing phase. This will help identify inefficiencies in the flow and guide changes to areas of concern.

As radiology research often involves multiple readers, this is also a good opportunity to ensure readers are similarly interpreting data elements when reading images. Research readings often involve different or more intricate levels of detail than standard clinical readings and take the radiologists outside of the daily reading routines with which they are most familiar.

A suggestion for performing a test run is to have study team members independently review the same small set of images (radiologists) or reports while filling out the DCFs (non-radiologists), followed by a discussion and comparison of the results. Principal investigators should be open to feedback provided during this stage and be willing to make changes to study materials or the process itself. If changes are made following IRB approval, an amendment may need to be submitted to the IRB to notify them about changes to the data collection process.

**Blinding**

If at all possible, those involved in data collection or reading images should be kept unaware of information that identifies important status or is strongly correlated with outcome. This reduces the potential for bias as the data collection or generation is unlikely to be affected by important characteristics of the study participants. Examples of information to be considered for withholding include diagnosis, prognosis, predisposing risk factors, and results of other tests or procedures. Blinding can often be accomplished by anonymizing images provided to readers so that radiologists do not have access to patient history or clinical reports, sorting images or records to be reviewed in random order for studies involving competing tests or re-reads, or using separate DCFs for different portions of the data collection process.

**Electronic Data Considerations**

Data will eventually be stored into a database or spreadsheet that is likely to follow a row-column format. Rows are intended to contain data for the units of measure (e.g., participants, lesions) and columns represent the variables themselves. It is advised to follow the row-column format, especially if working with a statistician, as this is the format that is recognized by analysis software packages.

**Standardizing Data Elements**

Data format should be consistent within columns. Common formats include dates, text (also known as string), and numeric. Only one data element or observation should be
included per column or cell. Deviations lead to confusion when reviewing the data and will likely involve laborious data cleaning to allow for the appropriate data analysis (Fig. 11.1).

Practical tips to aid with this include formatting the column so that it adds the appropriate number of decimal places or configures dates into the desired format. Many database or spreadsheet programs also allow data to be entered by selecting an option from a dropdown list or checkbox, which minimizes keystrokes and reduces error; these features are useful for entry of categorical data.

**Data Coding**

The preferable mode of data storage is as a numerical code. A simple example is gender, where female and male are two options. Arbitrarily, if possible, this variable could be entered as 1 for female and 2 for male. This will make it easier to perform data analysis, where numeric codes are often required for more advanced analytical techniques. Although this will require entry of numeric codes for basic spreadsheets, the use of database programs that include various data entry options might allow for built-in numeric coding. Continuing with the above example, such a program could allow for female to be selected from a dropdown list and be stored as a 1 in the actual data table that will be used for analysis or exported for use by a statistician. These and other more advanced database options are discussed below.

It is to be expected that a certain number of participants will have missing data. It will be useful to distinguish between items that are evaluated but intentionally left blank because the data could not be identified versus those that were left blank because they were unintentionally passed over by study personnel. Two important considerations when determining a code for missing data are to maintain the format of the variable in question (i.e., if numeric, the code should be numeric and not string) and to use values that are impossible for a given variable to indicate it is missing (e.g., 999 for age or 99 for body mass index).

Shading portions of the spreadsheet with different colors, while helpful to visually highlight important variables or subgroups, should not serve as a replacement for variables or proper coding as colors cannot be integrated into analysis.

**Handling of Personal Health Information**

The responsible use of personal health information (PHI) was discussed in detail in the ethics section, and much of what was described there applies to this section. On a

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Procedure date</th>
<th>Completed</th>
<th>Date of first visit</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30-Oct-12</td>
<td>Y</td>
<td>4-Jan-63</td>
<td>125</td>
</tr>
<tr>
<td>2</td>
<td>30-Sep-12</td>
<td>N</td>
<td>12/13/1981</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>12-Oct-12</td>
<td>No</td>
<td>Aug. 6, 1967</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>13-Oct-12</td>
<td>Y</td>
<td>17-Apr-68</td>
<td>165</td>
</tr>
<tr>
<td>5</td>
<td>4-Oct-12</td>
<td>Y</td>
<td>3-May-70</td>
<td>103</td>
</tr>
<tr>
<td>6</td>
<td>27-Oct-12</td>
<td>N</td>
<td>5-Jun-69</td>
<td>137</td>
</tr>
</tbody>
</table>

Fig. 11.1  Common errors with standardizing data elements. Examples of three issues that will cause problems during data analysis. The completed column contains both N and No responses that are intended to reflect the same thing. The date of first visit contains dates in three different formats. The weight column contains a ?; this entry is string while the variable is intended to be numeric.
more practical level, assigning a unique study identifier (study ID) for each participant on the DCF and in the database will go a long way toward reducing the need for entering some PHI parameters into the study database. If it is necessary to re-identify participants, a file that links study ID to PHI should be maintained separately outside of the database, and described in the original IRB application. Readers are encouraged to verify the regulations that govern the collection and use of PHI at their local institute.

Data Security and Storage

All study files containing data should be encrypted and only relevant study personnel should be provided with the password. If possible, such files should be stored on a secure network server, as the server will already have limited access and contents are likely backed up regularly. The use of a shared drive will also allow access to the data from multiple computers, which is a practical advantage for studies that involve multiple users or require work performed in multiple physical locations. Storing data on portable flash drives or the hard drives of computers has the potential for problems from lost or damaged hardware. Readers are encouraged to look into what computing resources are available to them that will allow for secure and efficient data storage. They should also clarify security and storage details with their IRB and describe measures that address these topics in their IRB application.

Multiple Observations per Participant

Some research scenarios are less commonly encountered in radiology than in other clinical research disciplines. They require special but simple considerations from study design and data management perspectives. The issue of multiple readers or techniques has already been discussed earlier in this section. Another important issue, multiple observations per participant, is described in further detail below.

The scenario where there can be more than one observation per participant is also known as clustered or correlated data. In radiology, outcomes or findings are not always limited to occurring once per person. Examples include brain lesions, lung nodules, and affected joints. If designing a study where individual findings are important, the details of each finding should be recorded separately. Put in terms used earlier, each finding should have its own row in the study database. Further, each level (finding, participant) should be considered a variable so that the combination of the two will identify what was observed in which participant (Fig. 11.2a). Collection of data in this manner will allow for the adjustment for correlation of observations within an individual during analysis. Standard analytical techniques assume independence of the observations, and this assumption is violated when dealing with clustered data. Specific analytical methods have been developed for clustered data,18,19 and they require identifiers at the level of the finding and participant. Improper accounting of these multiple observations (Fig. 11.2b) will lead to inaccurate results and incorrect conclusions. Further, the review process for peer-reviewed journals will usually involve individuals with a statistical background, and they are likely to flag studies in which data has not been analyzed appropriately.

Advanced Database Options

If certain technical or financial resources are available, some studies will benefit from the incorporation of more advanced database options. Some are available in menu-driven commercial programs that include the potential for customization with macros. Others will require they be incorporated by a programmer who will need clear direction from the study team to build a customized database. The following is a brief list of features to consider.

Electronic data entry forms are just what the name implies. Many database programs allow for the design and use of forms for data entry. Their use is easier and more intuitive than data entry directly into a spreadsheet. The layout of the forms can be designed to mimic the look of hard copy DCFs, which
### Fig. 11.2 Coding multiple observations per participant.

A proper coding scheme identifies participant (Patient ID) and observation (Lesion ID) so that combination of the two is unique (a). Details for more than one observation (Lesion ID) should not be made in a cell or row, as this will cause problems during data analysis (b).

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Lesion ID</th>
<th>Largest dimension</th>
<th>etc.</th>
<th>etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Largest dimension</th>
<th>etc.</th>
<th>etc.</th>
<th>etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4,4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6,4,3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
makes the entry process efficient and less prone to errors. DCF features like checkboxes, circles, etc. can be incorporated into the electronic forms. A feature of great benefit for analysis is the use of built-in numeric coding. Following the coding example discussed earlier, the data entry person can select female from a dropdown menu, and this can be stored as a 1 in the linked spreadsheet.

Web-based data are entered into a database that is stored on a network drive but can be accessed from multiple, even off-site computers via online login. This will of course require more advanced encryption. Before pursuing this option, interested readers should inquire with their local IRB whether a web-based database is allowed as this may be prohibited for security purposes. Such databases will usually also allow for simultaneous use and login by multiple users.

Finally, advanced security features should be considered for complex studies with larger study teams, especially when certain users do not need to have access to the entire database. These features can allow for each user to have a unique password, for access only to specific portions of the data, and for an audit trail of login details.

Quality Control

Frequent early data reviews are important to minimizing bias and maximizing accuracy. The identification and correction of errors early in the study are easier and less costly from a resource utilization perspective than if left until data collection has been completed. The identification of certain types of errors is relatively easy to incorporate into a quality control scheme by alerting study team members involved in data collection and entry to key study design features and objectives. These errors include extreme or out of range values (for dates and numeric variables), nonsense combination of responses (e.g., females with a history of prostate cancer), variables that have consistently missing data, and the inclusion of participants who should be excluded. While rules related to more formal internal data audits will depend on the design and nature of the study and be at the purview of the principal investigator, they nevertheless should be in place prior to study commencement, identify when and by whom audits are to be performed and who is to be notified if issues arise. If a data review results in changes to an approved study, the IRB may need to be notified.

Data Flow

The data management section has described numerous individual components of data management. A worthwhile exercise to better understand how data will flow is to view the study as a whole that is made up of these individual parts. Principal investigators should be aware that different people will be involved with different parts of the study and their activities must be coordinated. They should also be cognizant of study team members who are not exclusively dedicated to the study in order to ensure their availability at the appropriate time points.

Budget Formulation

The Merriam-Webster dictionary defines a budget as a statement of the financial position of an administration for a definite period of time based on estimates of expenditures during the period and proposals for financing them. The level of detail and complexity of a research proposal budget depends largely on the funding source and its requirements. Budgets can range from values as low as $20,000 in seed funding for pilot data, with rough-order of magnitude budget approximations and no reporting expectations on the funds usage, to several million dollar program grants with detailed bottom-up budgets with quarterly breakdown as well as specific milestone and deliverables reporting obligations. Regardless of the complexity, the main purpose of a research budget is to identify the essential costs associated with the proposal that should also be apparent in the proposal itself. The key questions to address while developing any budget include: Will the budget allow completion of the proposed project? Is the budget consistent with the activities of the proposal?
The specific contents, format, and layout of a research budget will vary based on the requirements of the funding body. General concepts applicable to many radiology-related budgets are described below.

**Expense Categories and Eligibility**

Funding sources can be divided into two main categories: internal sources where the funds originate at the applicant’s own institution and external sources, such as granting agencies, foundations, or industry partners. Expense categories applicable to most of these include personnel, equipment, materials, and supplies, broken down to varying level of detail, as demonstrated by templates in Fig. 11.3. Items that may be included are investigator salaries, institutional overhead, office equipment including computers, and travel expenses.

**Common Budget Omissions**

Items commonly missed from budgets include institutional overhead, pass-through costs (events invoiceable as they occur, e.g., an audit), staff benefits, research ethics board review fees, patient reimbursement fees, administrative fees associated with the use of various departmental services, annual salary increases, shipping fees, and data storage fees (clinical trial documentation must be stored for up to 25 years). Data transfer includes anonymization, CD/DVD burning, mailing, and electronic transfer and is often part of radiology research studies. Omission of budget line items will result in them not being reviewed by the funding agency. For successful funding recipients this may jeopardize the overall feasibility of the study, depending on the nature and value of the omission. Further, it is important to pay close attention to the allowable budget items and their specific restrictions as granting bodies may otherwise approve a lower amount submitted in the original application or reject the application.

Many funding sources allow for specific expense categories, such as statistical analysis or knowledge transfer activities and omission of these from a proposal and its budget may weaken the overall application. For example, lack of a statistician or an epidemiologist in the study team may be interpreted as a limitation in the ability to analyze the data or result in questions on appropriateness of sample size estimates.

**In-Kind Contributions**

In-kind contributions are nonmonetary resources that can be used to support the study. In a situation where total funding available

<table>
<thead>
<tr>
<th>Budget</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cash or in-kind contributions (direct research-related costs)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salaries (professional, technical, students, fellows)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equipment, software</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Material and supplies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge translation (travel, seminars, workshops, publication)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>In-kind contributions (overhead/indirect costs)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salaries of managerial and administrative staff</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of organization’s facilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grand total</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 11.3 Budget template. Example of a basic template with main items that should be considered for inclusion in a budget. The format and layout allow researchers to easily understand where costs will go and when they will be incurred during the study period (a). (continued)
Research Methods in Radiology

Fig. 11.3 (continued) Example of a more detailed template, commonly found in industry-sponsored clinical trials (b).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Cost/Unit</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informed consent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical history (including weight and height)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vital signs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Echocardiography status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum blood test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carcinoembryonic antigen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver volume/mass calculation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomize patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiogram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear medicine tech/prep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order exams/administer exams</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administer 2nd line chemotherapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Record/administer any chemotherapy following 2nd line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of life questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spiral computed tomography abdomen/pelvis/chest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverse events</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review/record concurrent medication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final endpoint efficacy/safety documentation &amp; exit patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study coordinator fees</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investigator assessments (including physical exam, tumour measurement, AE assessment and study oversight)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study patient travel expenses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-total for treatment group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overhead per patient (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL cost per patient in treatment group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-total for control group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overhead per patient (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL cost per patient in control group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Start-up fee</th>
<th>Standard charge</th>
<th>Overhead (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study start-up fee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imaging administration fees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archiving charts/storage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study close-out costs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total one time study costs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pass thru Invoice costs (% overhead)</th>
<th>Standard charge</th>
<th>Overhead (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen failures (maximum of 5 screen failures)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serious adverse event/case</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal or external audit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional work up &amp; administration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nursing costs per patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor measurements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitor unit per day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DVD burn per image request</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research Ethics Board (REB) fee</th>
<th>Standard charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>REB initial review</td>
<td></td>
</tr>
<tr>
<td>REB amendments</td>
<td></td>
</tr>
<tr>
<td>REB annual review</td>
<td></td>
</tr>
</tbody>
</table>
Obtaining Ethics Approval, Data Management, and Budget Formulation

Reporting on Milestones and Deliverables

Satisfactory reporting on milestones and deliverables identified in the original grant application is generally tied to the release of the next milestone payment and is used to ensure that the project is on budget and generally performing as outlined in the original grant submission. Reporting intervals range from quarterly to annual reports and commonly consist of milestone updates such as participant recruitment numbers (planned vs. actual), activities used to bring project back on track (if delayed), major scientific accomplishments, staffing changes, and a financial summary. In some cases, the payments may also be tied into the timing of the budgeted expense so that if travel was stated to occur in the second year of the funding, it would not qualify as an eligible expense in the first year.

■ Highlights of Key Points

Research Ethics Board Approval and Informed Consent

- The main purpose of regulatory guidelines and approval process is the protection of research participants, investigators, and institutions, with special emphasis on (1) protection of the rights and well-being of research participants and (2) the assurance of data integrity.
- Major regulatory documents include the Nuremberg code, Declaration of Helsinki, FDA CFR Title 21, parts 50 and 56, and, the most universally used document, the ICH GCP guideline.
- Research involving human participants must be reviewed and approved by an Institutional Review Board (IRB) prior to study initiation. Depending on the scope of the research study, additional approvals may be required by other authorities.
- Prior to initiating a research study, investigators must familiarize themselves with their local IRB regulations and the documentation required to obtain study approval.

Budget Justification

Budget justification in funding applications is critical as it provides an opportunity to answer any questions from the application reviewers. Radiology is a technology-dependent field which often requires high-costing budget items. It is thus advisable to include quotes from manufacturers or suppliers to demonstrate appropriateness of the pricing. For expensive service charges such as imaging, a pricing letter from the department chair offering the services is recommended.

A breakdown of line items to finer detail should also be part of justification. For example, rather than including a single line item for “imaging” at $21,250, itemization for $12,500 at $500/hr × 25 participants for 1-hour MRI scan, $2,500 at $100 × 25 participants for MRI contrast, and $6,250 at $250/scan × participants for CT scan, with written service letter from the institution providing the services, is suggested.

through the grant would not be sufficient to cover the cost of the entire study, in-kind contributions can be used to demonstrate that the study as a whole is feasible even if funds for these budget items are not provided by the grant being applied for. Depending on the source of funding, in-kind items may include time from staff that are already funded through other programs and who would be available to support the new proposal as well. Important examples are a study coordinator for participant recruitment or a radiologist to provide secondary reads. Examples of other in-kind contributions include pathology services through a collaborator, contrast agents or medical devices provided by industry sponsors, or imaging scan time from the investigator’s institution.

Academic centers and hospitals may waive some costs, such as research ethics board review fees, or apply reduced overhead charges to align with a specific grant limitation for academia-funded studies. The service providers should be consulted on these details prior to budget finalization to determine the level of institutional in-kind service and to ensure accurate budgeting.

Doria_CH11.indd   207
9/29/17   3:17 PM
Prior to carrying out any research procedures, consent must be obtained and properly documented, including a signed ICF by the participant and the consenter. The responsibility for keeping research-related data confidential lies with the principal investigator.

Data Management

- Data definitions should be objective, reproducible, and generalizable.
- A data dictionary is a document that lists and describes data elements. It is essential for data collection, entry, and analysis and should be made available to the study team.
- Data collection forms are useful tools with a template format that prompts users to record data in a consistent fashion.
- Test runs for data collection and entry may identify potential problems with data elements or inefficiencies that could be encountered by the study team.
- Data elements should be standardized for entry and coding.
- Databases should be encrypted and participants de-identified.
- If individual findings are of interest to the researchers, findings and participants will require separate identifiers to allow for the appropriate analysis of this correlated data.
- Audits that identify certain error types should be performed and, if possible, conducted early in the study.

Budget Formulation

- Questions to keep in mind while developing a budget:
  - Will this budget allow completion of the proposed project?
  - Is the budget consistent with the activities described in the proposal?
  - When designing a budget for a funding application, it is important to pay close attention to the eligible budget items and their specific restrictions. The application may be rejected if these are overlooked.
  - Items commonly missed from budgets include institutional overhead, staff benefits, Research Ethics Board review fees, patient reimbursement fees, service department administration fees, data storage, and transfer fees.
  - Budget justification is important. Quotes and support letters for expensive services or supplies are recommended.
  - Explore in-kind opportunities.
  - Prepare your budget in the context of reporting milestones and deliverables if continued funding is dependent on them.

References

13. Elger BS, Caplan AL. Consent and anonymization in research involving biobanks: differing terms and norms present serious barriers to an international framework. EMBO Rep 2006;7: 661–666
19. Lieber ML, Ashley C. McNemar extension: %Clust-pro macro. Cleveland, OH: Cleveland Clinic Foundation; 1998
respondents had decided on an academic career path during their second half of residency of radiology or later. The influences most often cited as leading toward an academic career were a desire to teach, the inspiration of a role model, and an interest in research. Therefore, when all is “said and done,” one would like to look back on one’s professional accomplishments and feel that he/she has been successful. Success means different things to different people and to the same person at different stages of his/her life or career. Success can be associated with a sense of achieving personal growth, having favorable relationships with others, maintaining a work/life balance, or attaining eminence or status. Some say “success” is similar to “happiness” in that “happiness is between the ears”—meaning that one is inherently happy or not and that no degree of material goods will affects one’s true happiness. However, for many individuals, success in their career will involve one doing meaningful daily work, being recognized nationally and internationally as a leader in their field, completing a body of work or research that directly or indirectly improves other peoples’ lives, being a role model for people entering the field, and being appropriately financially remunerated. Nevertheless, tips for professional success include obtaining early accomplishments, having highly successful mentors, and choosing coauthors who will actively contribute to the conduct of someone’s projects.

Prior studies have shown that one of the biggest barriers preventing recent graduates from entering academic radiology is the perceived pressure to write successful grant applications in order to be promoted. This may have been the case a generation ago, but is not the case in our current environment. There is much more to academic radiology

Learning Objectives

- To give useful information and advice to young faculty embarking on an academic radiology career. Navigating through the research process can be difficult and it is very important to start on the right footing.
- To provide anecdotal practical tips that cover academic faculty roles, getting started in research, grant writing, the mentoring process, and the necessary elements required to ensure continued research and career success.

Introduction

One of the first questions a trainee radiologist has to ask is, “Should I enter academics or private practice?” Historically, approximately 75% of radiologists in the United States enter private practice and about 25% choose academics. However, recent changes in the economy, including health care reforms that specifically affect radiologists, have resulted in a higher percentage of trainees entering academically oriented groups. The decision to enter a certain type of practice is not a straightforward one and will be impacted by various factors including family, spouse, financial status upon completion of residency, geographic choices, and drivers of success.

Most physicians consider medicine a “career” rather than a “job.” Concerning radiology, results from a survey conducted more than two decades ago in British Columbia, Canada, among graduate residents and staff radiologists with and without a university affiliation, pointed out the fact that most academic radiologists make a decision on academic radiology because of the teaching, research, and publication experience. Most
than writing grants. Studies have shown that among the reasons to pursue an academic career are the research opportunities available and the presence of positive role models. There are now four broad components of being a modern-day academic radiologist, which include (1) being an excellent specialized or subspecialized clinical radiologist, (2) teaching and scholarship, (3) research, and (4) administration/committee service. These four “legs” of the academic mission provide tremendous career latitude and expose those interested in pursuing an academic career to opportunities that were not previously available. In addition, the boundaries between so-called “academic” and “nonacademic” practices are now less distinct with many opportunities to excel available in larger “semi-academic, semiprivate” practices.

The majority of radiology residents will complete a subspecialty fellowship. Academic medical centers (AMC) are often tertiary or quaternary referral centers and require physicians who are subspecialized in their respective specialties. We tend to forget that the 99.9% of people in the world who are not physicians expect to get the best care if they choose to be cared for in an AMC. Therefore, it is essential that AMCs recruit and retain experts in their respective fields to meet their clinical mission. Thus, anyone who has completed a radiology fellowship meets the basic requirements to work in an AMC.

By definition, the majority of AMCs are affiliated with medical students, residents, and fellows who must be trained and educated. The majority of clinical training is performed by clinical faculty in largely an apprenticeship model. The primary mission of a medical school is education. One of the most substantial changes that have occurred over the past 20 years is the recognition that teaching is an essential role of a medical school, therefore it is imperative that medical schools reward faculty who are excellent clinician-educators through the promotion process. Historically, promotion was based on the “publish or perish” principle. However, the majority of AMCs now understand that there is nothing wrong with being an excellent clinician-educator and have created promotion tracks that allow faculty promotion based on excelling in these two important missions.

The surest way to a speedy promotion is through obtaining independent National Institutes of Health (NIH) or Canadian Institutes of Health Research (CIHR) or equivalent government-sponsored funding, particularly for faculty on the tenure or research track. This principle has not changed over the past 50 years. However, it must be emphasized that the majority of academic radiologists have never submitted a grant or been the recipient of an NIH R01 or equivalent research grant. There are many components to “research” that include writing case reports or series, performing prospective trials, writing book chapters, participating in educational or scientific exhibits, and performing retrospective case series, prospective studies, and clinical trials. The majority of academic radiologists participate in these clinically oriented studies as opposed to basic science, and this type of clinically oriented scholarly activity has been the basis for the majority of faculty promotions in academic radiology. Increasingly, there are opportunities to pursue research in other areas including evidence-based medicine, society standards, or clinical guideline development.

We formerly heard about the “tripartite” mission of academic medicine that refers to clinical work, teaching, and research. However, the unceasing regulatory scrutiny that has defined the last 15 years has changed our three-legged “stool” into a four-legged “table” with the fourth leg representing administration and committee work. There is a unique skill set associated with being a very good administrator that involves an aptitude to tackle never ending regulatory hurdles and to manage people. There is a science to administration and management, and these skills are part of standard business school curricula. However, these essential people management skills are not taught in medical school, and many physicians chose to study these in formal postgraduate education or have acquired them over time. Undoubtedly, the administrative aspects of health care are one of the largest growth opportunities in academic medicine.
Knowing that young physicians are embarking on a “profession” rather than a “job,” what can young academic physicians do to prepare themselves for a successful career in academic medicine? There is no specific answer to this question since there are no “large-scale, prospective, evidence-based studies” that have investigated the answer to this question. However, there are specific steps one can take to make an academic career fun, rewarding, and successful. We cover some of these opportunities in this chapter.

### The First Five Years

The scope of academic radiology continues to expand and provides a broad range of professional opportunities. Common questions that young faculty often ask include: “Where should I start?” “What should I get involved with?” and “Should I get involved in projects that I need to do or that I want to do?” There are no correct answers to these questions, but there are some general guidelines that may be useful. Unless you are absolutely certain about the specific area of academic radiology you wish to pursue, it is beneficial to get wide exposure to the field.

Most academic radiologists will already have subspecialty training, so the clinical aspect of the position is fairly straightforward. However, it is very important that you establish yourself as a good clinician as this will help with future multidisciplinary collaborations. Dedicating oneself to the clinical work when on service is important too, since being a good clinician will increase your reputation among clinical colleagues. In addition, the clinical work and patients encountered often generate the ideas for valid research questions or meaningful projects.

You will also be interpreting studies with trainees that include medical students, radiology (and clinical) residents, and fellows. As a teacher and consultant, you are acting as a role model to young trainees and imparting the basic and essential knowledge regarding how to be a clinical radiologist. You may also be their first image or “the face” of your specialty. Thus, you should try and establish a good rapport with these individuals and try and allocate at least 30 minutes each day for dedicated teaching when you are on the clinical service. An alternative if one is busy is to think aloud, or to impart at least one “teaching point” on each case that you interpret. An important lesson learned over time is that your ultimate reputation will be determined by those with whom you have interacted (your colleagues and your supervisors) and the individuals you train. We live in a very “small world” and both good and bad news travel fast. The medical field is even smaller and so your reputation will spread wide and far. Thus, it is important to foster good relationships with your trainees, allied medical colleagues, and administrative personnel. In addition, given the increased popularity of 360-degree physician performance assessments when it comes to retention, promotion, and tenure of faculty, this makes more than good common sense.12

New faculty come to AMCs with varying experiences with research. Most AMCs require some scholarly production for promotion. However, this requirement varies widely with the institution and the promotion track. Radiologists with little exposure to research should begin to work on small projects such as case reports, case series of unusual disease entities, or unique findings and educational exhibits. It is important to get engaged in research activities early as the first 5 years of activities have an influence on the research career trajectory later on and are a predictor of future research productivity. It can be very intimidating to get started on that first paper, and we highly recommended that one finds a mid-career or senior colleague who can mentor one through the process from start (idea generation or research question) to finish (manuscript drafting and publication). Young radiologists should try and identify colleagues in their own or other departments (e.g., clinical) with similar interests to work on collaborative projects and share responsibilities as this will dramatically increase scholarly productivity over time. Collaborating with clinical specialists will also increase one’s visibility among clinical colleagues, and when it comes to publishing, it opens up avenues...

---

Doria_CH12.indd 212
9/29/17 3:18 PM
for publication in high-impact mainstream journals. Junior radiologists may also wish to begin to review for journals. Many radiology journals are looking for reviewers and will assign manuscripts based on the complexity of the manuscript and experience of the reviewers. It is important not to review for too many since one can easily get overwhelmed. We began reviewing straightforward papers like case reports in our fellowship and junior faculty years and found that critiquing the writing of others greatly improved our scientific writing. For the initial journal reviews, it is a good idea to consult a senior colleague, who can read over the review and approve or edit it as appropriate and give general advice before submitting the review. The discussions that ensue will guide the young radiologist and build his/her reviewing skills as well as provide a networking opportunity with the senior colleague. From the academic institution’s perspective, if the junior radiologist receives some protected time for academic activities (20%, for example), he/she will be expected to produce scientifically as a priori determined by his/her department (e.g., to publish a minimum of one paper per year, to present papers and posters in scientific meetings, to participate in educational activities, etc.).

Some faculty will have prior research experience upon entering their first academic position and may wish to pursue extramural funding. As mentioned above, the surest guarantee for academic promotion is obtaining an NIH R01 or equivalent research grant. However, R01s are difficult to obtain, especially for junior faculty and subspecialists such as radiologists. An important consideration for these young investigators is when to apply for extramural funding. There have been numerous promising young radiologists whose first academic pursuit was to apply for an NIH/CIHR grant. This may not be the best strategy since the funding rate in 2012 was 18% and the average age at which a principal investigator receives his/her first R01 (NIH) is 42 years old (PhD applicants) and is even older for MD and MD-PhD applicants. Writing a grant is similar to writing a business plan for a venture capital firm. The funding agency wants to make sure they will receive a benefit for their investment. In the business world, this means a monetary return on investment. In the scientific world, the NIH/CIHR wants to make sure the proposal they fund is scientifically rigorous, the proposed project will be completed on time, and the results will be published in the peer-reviewed literature. The best predictor of future success is past performance and the NIH/CIHR is more likely to fund an investigator with a prior history of funding and relevant scientific publications as opposed to someone whose first grant submission is an NIH R01/CIHR Operating Grant application. There have been numerous investigators who spent the first few years of their academic career solely working on an R01/Operating Grant that was not funded. Consequently, these young investigators became discouraged and left academics because they only focused on their R01 applications and had no other scholarly achievements during this time period.

A better strategy for young faculty is to begin to write abstracts and small papers on a subject of interest. The next step is to apply for a small career development award that provides funding to acquire preliminary data. Tips on strategies to succeed as a young investigator can be obtained at specific forums or scientific meetings (Table 12.1). Getting started on that first grant proposal can be daunting and so there are many short courses on grant writing available, many of which will include the practice of writing an actual NIH/CIHR type grant proposal and receiving critique from colleagues and experts in the field. Asking someone to review your grant application ahead of time is key (ideally, two experts in the field and one person outside the field). We should keep in mind that some grant reviewers are not experts in the grant proposal field. Examples of grant writing courses and workshops on research proposal development specific to radiology are provided in Table 12.2. Career development awards are generally awarded to faculty within 5 years of completion of their residency or fellowship) training and funds can be awarded toward supplies, equipment, or scanner time or toward salary support of...
Table 12.1 Workshops on career development for young investigators

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARRS Clinician Educator Development Program</td>
<td>Nominees are selected and receive a grant to attend a 2-day workshop for junior faculty physicians to gain proficiency in teaching skills and designing educational activities. Interactive workshops are presented by leading educators and cover such topics as: adult learning theory and practice; developing curriculum; interactive techniques and exercises; case-based learning; designing a teaching-learning website; developing a teaching portfolio and writing self-assessment modules.</td>
<td><a href="http://www.arrs.org/RoentgenFund/Awards/CEDP.aspx">http://www.arrs.org/RoentgenFund/Awards/CEDP.aspx</a></td>
</tr>
<tr>
<td>AUR Academic Faculty Development Program</td>
<td>Candidates are nominated by the department chair of each participating university. This is a 1-day program, held during the AUR annual meeting, for junior radiology physician faculty members. The program is comprised of several presentations addressing such topics as: informatics in imaging; opportunities in education; how a chair can help your career; update of ABR, maintenance of certification; ethical issues and advice for publishing and peer review; funding opportunities through the RSNA; opportunities in research; and how to attain and maintain academic productivity.</td>
<td><a href="https://www.aur.org/AnnualMeeting/Faculty_Development_Program/">https://www.aur.org/AnnualMeeting/Faculty_Development_Program/</a></td>
</tr>
<tr>
<td>CIHR Young Investigators Forum</td>
<td>This forum is not focused on radiology, but aims to bring young investigators from different areas together to interact with one another, and attend workshops important for career development and related skills: Grant writing, time management, research team, translational research, etc.</td>
<td><a href="http://www.cihr-irsc.gc.ca/e/49888.html">http://www.cihr-irsc.gc.ca/e/49888.html</a></td>
</tr>
</tbody>
</table>

Abbreviations: ARRS, American Roentgen Ray Society; AUR, Association of University Radiologists; CIHR, Canadian Institutes of Health Research; RSNA, Radiological Society of North America.

Some junior radiologists without prior formal education in research may use part of this “given” time for research through salary awards to pursue a graduate degree, which should equip them with the methodologic tools required to succeed in future grant competitions and in a long-term academic career through the “research track.”

Given the health system reforms that are upon us, grant funding may in fact become more difficult, but on the other hand, there will be a greater emphasis on sponsoring projects that are patient centered or that look at health services aspects such as comparative effectiveness of imaging modalities. These areas offer exciting new avenues for focus of future research efforts. In addition, librarians, division directors, department directors, and mentors can help young academics identify funding opportunities and direct them to relevant organizations and funding sources.
Many investigators will start writing the next, larger grant proposal during the time of their initial career development award, using the preliminary data obtained during this time as a foundation to the larger grant project. Doing this will ensure a smooth transition to and continued uninterrupted protected time to devote to the larger body of research. Many of the oncology and other medical specialty societies also have grants available within their disciplines, which will include a substantial imaging component (Table 12.4). This is not surprising, given the impact that imaging has and the role it plays in clinical patient care and follow-up. This "stepwise" strategy will allow the investigator to develop a career focus and obtain both experience and preliminary data that will enhance their scholarly achievements and increase their likelihood of a successful NIH grant application.

Most senior academic radiologists agree that it is better to concentrate on the clinical, teaching, and research components of academic medicine during the first 5 years of your career. As a general principle, young faculty at the assistant professor rank should review the criteria of their promotional track and focus on the requirements to ascend to the next level (associate professor). The specific promotion requirements vary with the institution and track but most of them consider the five Ps for academic promotion: Publications, Principal investigator role in grants, Plenaries participation, Popularity, and Public service. Most institutions have a track for individuals who will focus on research and scientific publications and a clinical track that will reward individuals for excelling at the clinical and teaching missions. As a general rule, direct criteria for evaluation of faculty for academic performance in the tenure research track include peer-reviewed publications (number of publications and citations, and impact factor of journals), grants (source

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSNA Writing a Competitive Grant Proposal</td>
<td>A 1.5-day grant writing program for researchers in radiology, radiation oncology, nuclear medicine, and related sciences, interested in actively pursuing federal funding. This program focuses on developing strong specific aims and tools for getting started on the grant process. Participants are typically academic radiologists.</td>
<td><a href="https://www.rsna.org/Writing_A_Competitive_Grant_Proposal.aspx">https://www.rsna.org/Writing_A_Competitive_Grant_Proposal.aspx</a></td>
</tr>
<tr>
<td>RSNA advanced grant writing course</td>
<td>Participants are MD or PhD faculty in a radiology, radiation oncology, or nuclear medicine program. This four-session (four Fridays and Saturdays) course prepares participants to submit a quality grant application to the NIH, NSF, or other equivalent institution. Candidate needs to be sponsored by their institution chair.</td>
<td><a href="http://www.rsna.org/Advanced_Course_in_Grant_Writing.aspx">http://www.rsna.org/Advanced_Course_in_Grant_Writing.aspx</a></td>
</tr>
<tr>
<td>NIH Grantmanship Workshop</td>
<td>A half-day workshop at the RSNA meeting to provide participants with an introduction to writing a grant proposal for the NIH, specifically related to the biological sciences. Registration to be completed on the RSNA meeting website. There is a $35 fee to register.</td>
<td><a href="https://www.nimh.nih.gov/labs-at-nimh/scientific-director/office-of-fellowship-and-training/career-and-professional-development/grantmanship-workshop.shtml">https://www.nimh.nih.gov/labs-at-nimh/scientific-director/office-of-fellowship-and-training/career-and-professional-development/grantmanship-workshop.shtml</a></td>
</tr>
<tr>
<td>RSNA Clinical Trials Methodology Workshop</td>
<td>A 6.5-day workshop for faculty members and fellows in radiology, radiation oncology, and nuclear medicine academic departments, to assist in the development of a protocol for a clinical evaluation of imaging modalities, ready for inclusion in an application for external funding. The candidate needs to be sponsored by their institution chair.</td>
<td><a href="http://pubs.rsna.org/doi/abs/10.1148/radiol.2413060113">http://pubs.rsna.org/doi/abs/10.1148/radiol.2413060113</a></td>
</tr>
</tbody>
</table>

Abbreviations: NIH, National Institutes of Health; NSF, National Science Foundation; RSNA, Radiological Society of North America.
Table 12.3  Career development awards for radiology faculty within 5 years of completion of their training

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Link</th>
</tr>
</thead>
</table>
| RSNA Research Seed Grant*                 | To enable investigators throughout the world to gain experience in defining objectives and testing hypotheses in preparation for major grant applications to corporations, foundations and governmental agencies. Up to $40,000 for a 1-year project to support the preliminary or pilot phase of scientific projects.  
* Open to international applicants | https://www.rsna.org/Research_Seed_Grant.aspx                                                                                                                                                           |
| RSNA Research Scholar Grant               | To support junior faculty members who have completed the conventional resident/fellowship training programs but have not yet been recognized as independent investigators. $75,000 annually for 2 years ($150,000 total) to be used as salary support for the scholar. Requires a minimum of 40% effort. | http://www.rsna.org/Research_Scholar_Grant.aspx                        |
| RSNA/AUR/APDR/SCARD Radiology Education Research Development Grant* | To encourage innovation and improvement in health sciences education by providing research opportunities to individuals in pursuit of advancing the science of radiology education. Up to $10,000 for a 1-year project.  
* Open to international applicants | http://www.rsna.org/Radiology_Education_Research_Development_Grant.aspx                                                        |
| RSNA Educational Scholar Grant*           | To provide funding opportunities for individuals with an active interest in radiologic education. One-year grant of up to $75,000 for salary support and/or other project costs. In exceptional cases, grants for up to 2 years will be considered.  
* Open to international applicants | http://www.rsna.org/Education_Scholar_Grant.aspx                                                                               |
| GERRAF Award                              | $70,000 for each of 2 years, paid through the sponsoring institution to be used primarily for salary support. Up to $10,000 each year may be used to secure assistance in executing the project. Additional $10,000 of the stipend can be used for coursework in an MPH or other degree program approved by the GERRAF Board of Review over the 2-year fellowship. Otherwise, the entire amount should be used exclusively for the fellow’s salary support. In return, the institution must make available not less than one-half of the fellow’s time to pursue the approved research and educational program. | http://www.aur.org/Secondary.aspx?id=111                               |
| ARRS Scholar Award                        | The ARRS Research Committee may select up to two scholars. A maximum of two scholarships of $140,000 are funded. Scholars may choose a 1-year program requiring a minimum 80% time commitment, or a 2-year program requiring a minimum 50% time commitment. The funds may be used for salary support, toward the support of the scholar’s study or in a way that will contribute to the scholar’s development and advancement as an academic faculty member. The money may not be spent for the purchase of equipment. Indirect costs may not be paid from this fund. | http://www.arrs.org/ARRSLIVE/ASNRScholarship                        |
| NIH K awards                              | Mentored Clinical Scientist Developmental Program Award (K12) award to specific institutions. Individual awards include the Mentored Patient-Oriented Research Career Development Award (K23). If you have already been trained and want to serve as a mentor to more junior clinicians, try the MidCareer Investigator in Patient-Oriented Research Award (K24). There are other awards that should be examined including the Academic Career Award (K07), and the Mentored Quantitative Research Career Development Award (K25). | http://grants.nih.gov/training/careerdevelopmentawards.htm            |
of funding, industry investigator-initiated vs. peer-reviewed national competitions, the latter one being more valued, amount and continuity of funding), mentoring and training of high-qualified personnel, and patents. Indirect criteria for faculty academic performance after the first 5 years of an academic career include participation in professional organizations (continuing medical education [CME] courses, committee positions, board and officer positions), journal-related activities (ad doc reviewer, editorial board, editor), services on national or international policy committees, and invitations as visiting professor or plenary speaker.\textsuperscript{19} For example, in prestigious universities in the United States and Canada, the research track promotion dossier of a candidate to assistant professor should demonstrate local reputation and some competence as an independent investigator. Specific investigator’s outcomes at this point of his/her career should include at least two first authored papers, invitations to speak at scientific meetings, ad hoc reviewer for journals, and role of co-investigator on grants. The research track promotion dossier of a candidate to associate professor should demonstrate national reputation of the applicant. It is expected to include a total of 25–50 peer-reviewed publications (one-third as first author, some as senior author), experience on mentoring/training of students (at least 10 prior students, at the undergraduate, graduate, or postdoctoral levels), grant funding as principal investigator (one past CIHR/R01 NIH grant is ideal, but other funding sources are acceptable), and editorial journal board and committees of professional societies participation (advisable). Finally, the research track promotion dossier of a candidate to full professor should demonstrate international reputation of the applicant. It is expected to include a total of 80–120 papers (most of them as senior author), substantial grant funding (2–5 previous R01 NIH/CIHR Operating Grants are ideal; nevertheless, proof of continuity of

Table 12.3  Career development awards for radiology faculty within 5 years of completion of their training (Continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIHR Young Investigator Award</td>
<td>In Canada salary of investigators cannot be requested in grant applications and protected time for research must be provided by the investigator’s institution. This award aims to fund academic institutions with salaries for highly skilled and qualified junior staff (within 5 years of first staff appointment). The evaluation of the New Investigator Salary Award applications is based on the following three criteria: (1) environment and support; (2) research plan; and (3) track record of the candidate. Successful applicants receive up to 5 years of funding as a salary ($60,000 per year).</td>
<td><a href="http://www.cihr-irsc.gc.ca/e/41208.html">http://www.cihr-irsc.gc.ca/e/41208.html</a></td>
</tr>
<tr>
<td>CCHCSP Career Development Award</td>
<td>This award provides support for highly qualified child health clinician candidates to develop their requisite knowledge and skills for a career as an independent scientist in child health research. Trainees engage in research training and a core curriculum in one of the participating Child Health Research Training Centres of the Canadian Child Health Clinician Scientist Program. Research training is supervised by a research mentor and advisory committee and structured to facilitate excellence in research with an interdisciplinary focus. Successful applicants receive up to 4 years of funding as a salary ($70,000 per year; 50% from CCHCSP and 50% from the home institution).</td>
<td><a href="http://cchcsp.ca/index.php/training-paths/training-path-details?intPath=3">http://cchcsp.ca/index.php/training-paths/training-path-details?intPath=3</a></td>
</tr>
</tbody>
</table>

Abbreviations: APDR, Association of Program Directors in Radiology; ARRS, American Roentgen Ray Society; AUR, Association of University Radiologists; CCHCSP, Canadian Child Health Clinician-Scientist Program; CIHR, Canadian Institutes of Health Research; GERRAF, General Electric Radiology Research Academic Fellow; NIH, National Institutes of Health; RSNA, Radiological Society of North America; SCARD, Society of Chairs of Academic Radiology Departments.
peer-reviewed funding is key), participation in previous NIH/CIHR study sessions, and board and officer positions at national and international professional organizations. Young faculty should familiarize themselves with the promotion requirements in their own institutions, talk with individuals who have recently gone through the promotion process, and discuss criteria with current or former members of the promotion committees.

Most institutions have an assistant or associate dean for faculty affairs whose responsibility is to counsel faculty on the promotion process. It is a good idea to have a mid-cycle review, say after 3 years, in which the faculty meets with the assistant or associate dean for faculty affairs to discuss their trajectory so far and to ensure that they are on track for their next promotion. Administrative responsibilities such as serving on hospital or department committees are considered by the promotions committee, but tend not to be as heavily weighted compared to the other missions, particularly for junior faculty promotion. Thus, it is best to avoid too many administrative responsibilities with large time commitments unless it is clearly associated with one of the three primary missions. A good example

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH R01</td>
<td>R01s can be investigator-initiated or can be in response to a program announcement or request for application. This website is devoted to the investigator-initiated R01 application, which means there are no specific program requirements. However, the R01 research plan proposed by the applicant must be related to the stated program interests of one or more of the NIH institutes and centers based on descriptions of their programs. R01s are most often investigator initiated in response to either the R01 Parent Announcement or a Program Announcement highlighting particular scientific areas. Requests for Applications (RFAs) may also utilize the R01 mechanism. Applications are generally awarded for one to five budget periods, each normally 12 months in duration.</td>
<td><a href="http://grants.nih.gov/grants/funding/r01.htm">http://grants.nih.gov/grants/funding/r01.htm</a></td>
</tr>
<tr>
<td>PCORI</td>
<td>A variety of funding opportunities designed to advance patient-centered comparative clinical effectiveness research (CER) in accordance with the National Priorities for Research and Research Agenda. Five priorities: Assessment of Prevention, Diagnosis, and Treatment Options; Improving Healthcare Systems; Communication and Dissemination Research; Addressing Disparities; and Improving Methods for Conducting Patient-Centered Outcomes Research.</td>
<td><a href="http://www.pcori.org/funding-opportunities/">http://www.pcori.org/funding-opportunities/</a></td>
</tr>
<tr>
<td>AHRQ</td>
<td>The Agency for Healthcare Research and Quality offers various opportunities for funding in areas including health care research and recovery act.</td>
<td><a href="http://www.ahrq.gov/funding/index.html">http://www.ahrq.gov/funding/index.html</a></td>
</tr>
<tr>
<td>NIBIB</td>
<td>National Institute of Biomedical Imaging and Bioengineering offers various funding opportunities related to medical technologies, with awards ranging from 2 to 5 years.</td>
<td><a href="https://www.nibib.nih.gov/research-funding">https://www.nibib.nih.gov/research-funding</a></td>
</tr>
<tr>
<td>ARRS Leonard Berlin Scholar-ship in Medical Professionalism</td>
<td>Scholars may choose either a 1-year program requiring a minimum of 50% time commitment to be funded at $100,000, or a 2-year program requiring a minimum 25% time commitment to be funded at $50,000 each year. Up to $20,000 of the funds may be used in a way that will contribute to the scholar’s development and professional advancement, such as tuition and course materials, with the remainder to be used toward salary support. The money may not be spent for the purchase of equipment, or imaging studies. Indirect costs may not be paid from this fund.</td>
<td><a href="http://www.arrs.org/ARRSLIVE/BerlinScholarship">http://www.arrs.org/ARRSLIVE/BerlinScholarship</a></td>
</tr>
</tbody>
</table>

Abbreviations: AHRQ, Agency for Healthcare Research and Quality; ARRS, American Roentgen Ray Society; AUR, Association of University Radiologists; NIBIB, National Institute of Biomedical Imaging and Bioengineering; NIH, National Institutes of Health; PCORI, Patient Centered Outcomes Research Institute.
protected time for application eligibility. Currently there is a low demand of available positions of Clinician-Scientists in radiology in North America, which is also a widespread problem in academic institutions around the world. Nevertheless, Clinician-Scientists play an important role in health research and in clinical practice since they are able to develop research questions based on clinical issues they encounter in practice and are able to rapidly translate their research results into clinical practice.20,21,22 Academic departments are responsible for the great majority of training and technical innovation in the specialty.23 Therefore, many academic radiology departments have as a priority the sustainability of their academic missions in which training of future leaders in the field is a key goal.

■ Mentor versus Supervisor

An academic career will be greatly aided by having a good mentor. The challenge is finding a good mentor and distinguishing between a mentor and a supervisor. A supervisor is an individual who can provide information, assign projects, and oversee their completion through a product-oriented process. A supervisor can facilitate a career by informing individuals they supervise of projects and other career development opportunities. However, the supervisor is generally a conveyor of “information.”

The term “mentor” originated in Greek mythology and has been adopted into the English vocabulary as someone who imparts knowledge and understanding to a less experienced colleague, or protégé.24 In contrast to “supervision,” which is a product-oriented process, “mentorship” is a person-oriented process. The responsibilities of a mentor may merge with those of a supervisor, thus including both assigning projects and informing the mentee of career development opportunities through a person-oriented process. Starting with basic projects that can be completed in a short time frame and then progressively moving on to more complex projects as the protégé becomes comfortable with the details is a good idea. A good mentor takes personal
Research Methods in Radiology

...production such as a scientific presentation or poster and eventually end up in a scholarly publication that will help with academic promotion. It is important to make the junior faculty feel like a vital part of the research group. This is more likely to stimulate them to stay in academics and pursue a research career. Good mentors know that with junior faculty, they often spend as much or more work helping design the project, monitoring the progress than assisting with the presentation and paper. However, they are still willing to allow the trainee to be the first author while they have enough experience to be the last (senior) author. Within a few years, the mentor will have the satisfaction of seeing the junior faculty blossom into a fully fledged researcher and writer. A common pitfall about having a mentor is that the protégé may feel that the mentor is not doing enough to support their career, which may be the case in some instances. This can stress the relationship and cause both parties to be unhappy. However, it is very important that the trainee appreciates the effort it takes to be a good mentor and try to ensure that the mentor gains something in the relationship beside personnel and professional satisfaction. Gaining an insight to this interdependency and dual transference will help strengthen the bond, resulting in a very successful mentor–protégé relationship.

Not all mentor–protégé relationships work and it is important to establish ground rules for ending the relationship upfront so that things end as amicably as possible. Even for mentor–protégé relationships that work, a time will come for the relationship to end, so it is a good idea to know the ground rules for separation in advance. It is also important for the junior staff to become independent from their supervisor at some point of their early career. Demonstration of this is to become senior author of papers conducted by previous trainee and supervisor where the previous supervisor becomes a coauthor.

### Time Management

Discussions about time management should be part of the agenda of meetings between...
junior radiologists and mentors. Covey’s seven habits of highly effective people rely on focusing on quadrant 2 of a four-piece circle (where quadrant 1 represents things that are important and urgent; quadrant 2, things that are important but not urgent; quadrant 3, things that are urgent but not important; and quadrant 4, things that are not urgent and not important). According to this planning circle, highly effective professionals would focus on planning, preparation (of manuscripts, grants, presentations, etc.), prevention, studying, and network building, without leaving aside urgent and important tasks such as clinical duties, deadlines, solution for crises, and reviews of galley proofs of papers and chapters.

Most successful academic physicians ruthlessly protect their time from interruptions during their academic day(s). A suggestion to achieve this goal is to group meetings and writing periods into “blocks” (mornings or afternoons) during academic days. Time management skills are especially important for radiologists who need to juggle between academic and household duties with families. Junior radiologists who are new to a given institution should discuss personal issues that may interfere with their academic productivity or professional time scheduling with their department chairs or research directors to avoid future disappointments. During these meetings the junior radiologist should propose potential solutions to overcome conflicts of interest.

Leadership

The characteristics associated with leadership are personnel attributes as opposed to professional titles. The top four characteristics associated with successful leaderships are honesty, forward-looking, inspiring, and competence. A leader is often defined by the group that an individual leads. As individuals we all have the potential to be leaders. Remember that the majority of the population does not have an advanced degree. Society looks to those who have attained advanced professional degrees such as doctors, lawyers, and successful business people as community leaders and role models. We hold a similar implicit standing with those with whom we work including residents, fellows, medical students, nurses, technologists, etc. Thus, we are all playing the leadership role and need to embody these characteristics as we determine the culture and tone in both our professional and personnel environment.

A common question is when to pursue leadership opportunities and when to accept such positions. Being an effective leader requires being a good communicator and acquiring the respect of those we lead. The latter quality is usually obtained by experience and by demonstrating that you understand the organization and the people and can offer “added value” to those who you lead. Embodying the principles of good leadership means living them out in your everyday life. Thus, it is probably better to focus on the clinical, educational, and research missions early in one’s medical career and use this time to acquire experience and cultivate the relationships that will lead to future leadership positions.

Leadership skills in research involve setting a good example by using the best evidence from the literature and experience to generate ideas, applying the highest ethical values and principles to all aspects of the project (patient recruitment, blinding, allocation, data entry, data analysis, manuscript drafting, and interactions with collaborators), meeting regularly with your coinvestigators and communicating openly, delegating appropriately to junior colleagues, and showing commitment by working hard. Later on, the young radiologist will learn to play a more senior or leadership role in projects with the junior or less experienced collaborators.

Most organizations and radiology societies offer various types of leadership development programs. These programs range from a few hours to several days with very variable curriculum. Table 12.5 displays some available leadership opportunities within radiology. These can provide a good introduction to leadership and will introduce the participant to the concept of personnel growth and developing traits that will prepare them for future leadership opportunities. If one desires
Ultimately, it is up to the individuals to decide if they wish to accept their flaws and make the necessary changes. So, only pursue an executive coach if you are ready to accept criticism and are willing to adapt.

In summary, a career in academic radiology has never been as exciting as it is at present. There are so many opportunities available nowadays for junior faculty to carve out their own individual career, whether this concentrates on research, teaching, administration, or clinical work or a combination of all of these. Peers, colleagues, and senior faculty are a great resource for information and mentoring and they should be consulted freely and often. Increasingly, there are grant funding opportunities for clinician educators and health services research. Having a good mentor or mentors is essential at all stages of one’s career.

### Highlights of Key Points

- Learn from role models, the good, the bad, and ugly.
- Talk to lots of people to hear their perspective—ask them what they did right,
what they didn't do right, and what they would do if they were in your shoes.

• If you want to get promoted fast, write a grant.
• Start with a smaller, career development grant, before applying for a larger, R01 type grant. Collaborate within your division, collaborate outside your division, collaborate outside your specialty, and collaborate outside your institution!
• Get a mentor or mentors, including from specialties and departments outside your own, for different aspects and stages of your career.
• Take a leadership development course to learn what motivates you and what kind of a leader you might be.

Thoughts

Ingredients for academic success19:

1. Persistence, enthusiasm, and optimism
2. Honesty and integrity
3. Creativity and a lack of fear of holding the minority opinion
4. Ability to be a team player
5. Taking time to mentor
6. Generosity of time and ideas
7. Open and direct communication
8. Humility, courtesy, and respect
9. Intellectual curiosity
10. Intelligence and knowledge

“It is amazing what you can accomplish if you do not care who gets the credit.” –Harry Truman (1884–1972)


“A pessimist sees a difficulty in every opportunity. A positivist sees an opportunity in every difficulty.” –Winston Churchill (1874–1965)

“Rejection is part of the game in research. Researchers are very robust people (like business people) since they must deal with rejection most of the time.” –Unknown thinker


“A leader takes people where they want to go. A great leader takes people where they don’t necessarily want to go, but ought to be.” –Rosalynn Carter (1927–present)

“If your actions inspire others to dream more, learn more, do more and become more, you are a leader.” –John Quincy Adams (1767–1848)

“A leader never deflects personal responsibility, never hesitates to implement difficult decisions or to review decisions.” –George H.W. Bush (1924–present)

“If two people agree on everything, you may be sure that only one of them is doing the thinking.” –Lyndon B. Johnson (1908–1973)

References

13 Statistical Inference: Point Estimation, Confidence Intervals, and Hypothesis Testing

Christopher Meaney, Rahim Moineddin, and George Tomlinson

Learning Objectives

- To explain point estimation.
- To describe interval estimation and confidence intervals.
- To describe the logic of hypothesis testing.
- To present methods of analysis of some common study designs.

Introduction

The previous chapter on exploratory data analysis and descriptive statistics (Chapter 2) covered methods for using an observed sample of data to describe and explore its most interesting features. However, we do not usually want to restrict our conclusions to the particular sample that we took, but instead, we want to use the data to make more general statements about relationships in the broader population. For example, if we take a sample of 20 patients with known liver malignancies and find that 18 of them had abnormal ultrasound findings in the month before their biopsy, what is the reason that we use $18/20 = 90\%$ as the estimate of the detection percentage for ultrasound in the population of patients with liver malignancies? Given this estimated value of 90\%, what are the smallest and largest plausible values for the detection percentage in the larger population? How sure are we that the actual percentage in all patients was not 60\% and we got “lucky” with this sample? This chapter explains the statistical methods for answering these questions.

To move beyond simple descriptive statistics, and to use a sample to make inferences about the population, we need to consider two key elements: (1) the statistical model which is presumed to have generated the observed data; and (2) the manner in which the sample was taken. The first element is the focus of this chapter; the second element, selection of the sample, is crucial to analysis of data and can have an influence on how generalizable the results of an analysis are to the wider population (see Chapter 5), but is not considered in any detail here.

Example of a Statistical Model

A statistical model is a construction that links unobserved (and unobservable) quantities in the population to the observed data. We will illustrate by constructing a model for the example introduced above, where the observed data are the abnormal ultrasounds in patients with liver malignancies and the unobserved quantity is the true detection percentage. First, we define the population as all patients with liver malignancies who have an ultrasound in the month before a definitive biopsy. Then, we introduce the idea of a probability, in this case the fraction of the population that has an abnormal ultrasound in the month before their biopsy. In statistics, this probability is usually represented by the letter $p$. To simplify the wording, we will refer to an abnormal ultrasound as an “event.” Finally, we have the notion of a sample, the number of patients (in this example) taken from the larger population and assessed. If we randomly sample one patient from the entire population, what is the probability that this patient will have an event? By definition, it is $p$ and the probability of the patient not having an event is $1 - p$. 
But what if we randomly sample two patients? Given that we observe a sample on two patients, what is the probability that we observe zero events on these two patients? One event? Two events? When the outcomes on the two patients are independent, we multiply the probability of the outcome on one patient by the probability of the outcome on the second patient to obtain the total probability for the two patients. So that we do not have to keep repeating the phrase “the number of events,” we use \( x \) to represent this idea. The probability that \( x = 2 \) is found by multiplying the probability that the first patient has an event times the probability that the second patient does:

\[
P(x = 2) = p \times p = p^2
\]

Similarly, the probability that neither patient has an event (so that \( x = 0 \)) is

\[
P(x = 0) = (1 - p) \times (1 - p) = (1 - p)^2
\]

There are two equally likely combinations of outcomes that give \( x = 1 \): the first patient has an event and the second does not; the first patient does not have an event but the second does. We need to add probabilities for both combinations to get the probability that \( x = 1 \):

\[
P(x = 1) = p \times (1 - p) + (1 - p) \times p = 2 \times p \times (1 - p)
\]

These equations can be expressed as a single equation that gives the probability of observing \( k = 0, 1, \) or 2 events in a sample size of 2. The notation \( \binom{2}{k} \) is read as “2 choose \( k \)” and says how many different ways there are to pick \( k \) items from a set of 2.

\[
P(x = k) = \binom{2}{k} p^k \times (1 - p)^{2 - k}, \quad k = 0, 1, 2
\]

If we now replace 2 with any value of \( n \), we have an equation that defines what is called the binomial model. It computes the probability of observing \( k \) events in \( n \) subjects, when each subject has the same probability \( p \) of having an event:

\[
P(x = k) = \binom{n}{k} p^k \times (1 - p)^{n - k}, \quad k = 0, 1, 2, \ldots, n
\]

**Estimating \( p \)**

How does all of this mathematics help us estimate the true value of \( p \) from sample data? The model provides the link between the data (e.g., 18/20) and the quantity we are interested in (\( p \), the probability of an abnormal ultrasound). Filling in \( k = 18 \) and \( n = 20 \), Equation 1 has only one unknown, \( p \):

\[
P(x = 18) = \binom{20}{18} p^{18} \times (1 - p)^{20 - 18} = 190 p^{18} \times (1 - p)^2
\]

A widely used method, maximum likelihood, uses this equation to find the value of \( p \) that makes the observed value of 18 most likely; in other words, it finds \( p \) at the maximum of the curve in Fig. 13.1, a plot of \( p(x = 18) \) versus values of \( p \) between 0.6 and 1. Among all possible values of \( p \), it is the value 0.9 that makes the observed 18 events in 20 patients most likely and this is what we use as an estimate, \( \hat{p} = \frac{18}{20} = 0.9 \); the “hat” or circumflex in \( \hat{p} \) ("p hat") indicates that it is not the true value (which we cannot know) but an estimate of it. It is possible to use calculus to find the value of \( p \) at the maximum of the curve in Fig. 13.1, but we don’t need to. The estimator in the binomial model is always \( \hat{p} = \frac{x}{n} \).

**Estimating a Confidence Interval for \( p \)**

Fig. 13.1 also shows us something else important—there are some values of \( p \) that are almost as good as 0.9 (e.g., 0.85) but there are some values (e.g., \( p = 0.6 \)) that have much less support from the data. Maximum likelihood theory provides a way to use the observed data (18 out of 20) and the model (binomial) to obtain a range of values for \( p \)—an interval—that has some degree of plausibility and to exclude from this interval values that are implausible. Most commonly, this interval is constructed to have 95% “confidence.” The technical definition of the 95% confidence interval is rather wordy: a confidence interval has 95% coverage if in 95% of repeated random samples from the same population, the interval that is constructed includes the true (unknown) value. Our advice is to remember this definition but to think of the 95% confidence interval as a
range of values that is very likely to include the true value. With 18 events in 20 patients; one approach to construction of a 95% confidence interval gives [0.72, 0.98]. Values for \( p \) outside this range are less likely than values in it, and we have 95% confidence that this interval covers the true value.

**Sampling Variability**

Fig. 13.2 shows the probability of each possible number of abnormal ultrasounds if the true probability of any one being abnormal is 0.7. These correspond to estimated \( \hat{p} \) values of 0/20, 1/20, 2/20, …, 20/20. Our estimator \( x/n \) has what is called a sampling distribution; if we know \( p \), we know the probability of getting each value \( \hat{p} \). The variability of \( \hat{p} \) around its expected value of 0.7 can be measured by a quantity called the standard error. For the binomial model, the standard error is equal to \( \sqrt{p \times (1 - p) / n} \). Fig. 13.3 shows the results of the same experiment run on a larger sample of \( n = 200 \), again assuming that the true percentage abnormal is 0.7; here, the standard error is \( \sqrt{0.7 \times (1 - 0.7) / 200} = 0.0324 \). Many estimators presented later in this chapter share a property of the estimate of the binomial probability: in repeated samples from the same population, they will approximately have what is called a normal distribution with a mean equal to the true value and variability about that mean equal to the standard error.

**Hypothesis Testing**

Suppose that the data in our example were collected in a radiology department that was attempting to achieve a detection percentage above 70%. With 18 abnormalities detected in 20 patients, can we conclude that this target is being met? To answer this question, we can set up two competing hypotheses, called the null \( (H_0) \) and alternative \( (H_a) \).

- \( H_0: p \leq 0.70 \) [The department is not meeting its goals]
- \( H_a: p > 0.70 \) [The department is meeting its goals]
Although our observed value of $18/20 = 0.9$ is clearly above 0.7, these hypotheses concern the true value. Only one of these hypotheses can be true and we need to make a decision, based on our observed data, about which one it is.

We start by noting that if we decided the target was met based on 18 events in 20 patients, we would also have made that same decision if we had seen 19 or 20 events on these 20 patients. If 18 is a large enough number that we conclude $p > 0.70$, then clearly 19 and 20 are also high enough. Thus, we can phrase the question as, “If the department is not meeting its goals, what is the chance of seeing ultrasound performance at least as good as I have observed?” or “If the true detection percentage is 70%, what is the probability that I will observe 18, 19, or 20 detected abnormalities in a sample of 20 patients?” Our statistical model (i.e., the binomial probability model) provides the answer to that question. If we assume that the true value of $p$ is as low as 0.7, we can calculate the chance of observing values as high as we did through this formula:

$$P(X \geq 18) = P(X = 18) + P(X = 19) + P(X = 20)$$

$$= \binom{20}{18} 0.7^{18} 0.3^2 + \binom{20}{19} 0.7^{19} 0.3^1 + \binom{20}{20} 0.7^{20} 0.3^0$$

$$= 0.0355$$

There is only a 3.6% probability of observing such high numbers (18, 19, or 20) if the detection rate is an unacceptable 70%. The bars for 18, 19, and 20 abnormalities, shaded red in Fig. 13.2, have a total probability of 3.6%. This number is called the $p$ value for a test of the hypothesis that the detection rate was 0.7 or less. But what should be done with the $p$ value? Does the radiology department decide that detection rates are not unacceptably low?

The $p$ value was introduced by the statistician R.A. Fisher as a measure of evidence against the null hypothesis: the smaller the $p$ value, the less likely the observed data under the null hypothesis and, therefore, the stronger the evidence against that null hypothesis. Notably, Fisher did not suggest that the $p$ value alone should be used to interpret the evidence but believed that it was one piece of a process that also took into account background information. According to this view, the $p$ value of 0.036 has to be interpreted by the investigator in light of other pertinent information. A 2016 consensus statement issued by the American Statistical Association, the primary American professional association for statisticians, emphasized this view, stating “Scientific conclusions and business or policy decisions should not be based only on whether a $p$-value passes a specific threshold.”

Hypothesis testing plays a central role in analyses of data from medical research, so we expand here on the logic of this approach. In general, there are four important elements to a statistical test:

1. The null hypothesis ($H_0$)
2. The alternative hypothesis ($H_a$)
3. The test statistic
4. The rejection region or the threshold value for rejecting $H_0$

Let’s return to our ultrasound detection rate example, where the head of radiology needs to make a decision as to whether the detection percentage is above or below 70%. We formulate our statistical hypothesis testing problem as:

$$H_0: \ p \ < \ 0.7$$

$$H_a: \ p \ \geq \ 0.7$$

The test statistic in this example is the number of the 20 cases that are detected. We use this statistic to decide whether to reject $H_0$. The rejection region will be a prespecified set of values (e.g., 15 or more; or 18 or more) that will lead us to reject the null hypothesis. The magnitude our test statistic forces us to make a binary decision as to whether or not we are in the rejection region, once we have observed the study data. The pairing of the null and alternative hypotheses with the decision to reject the null hypothesis or not reject it means that there are four possible outcomes of a hypothesis test. These are shown in Table 13.1 where, to make matters concrete, we have set the rejection region to be either 19 or 20 abnormal ultrasounds.
We observed $x = 18$, so with our rejection region, we will not reject $H_0$. We are not able to reject the hypothesis of a low true detection rate. Importantly, we do not accept the null hypothesis and decide that the detection percentage is below 70%; that would be bizarre, given that we observed a detection percentage of 90%. The conclusion is that the count was not high enough to rule out a true value of 70%. We are in the second row of Table 13.1, where we are either correct to not reject $H_0$ (column 1) or we are incorrect (column 2); we don’t know which column we are in, because we don’t know whether the null hypothesis is true or not. But if in fact the null hypothesis is false, when we do not reject it, we commit what is called a type 2 error.

Now let’s assume that the head of radiology demands a second study assessing the liver abnormality detection percentage on ultrasound. In this second study, analyzed according to the same rules as detailed in Table 13.1, 19 of 20 cases with liver malignancies had an abnormal ultrasound. In this second study, we reject the null hypothesis, so we are situated in the first row of Table 13.1 and we are either correct to reject $H_0$ (column 2) or we are incorrect (column 1). If in fact the null hypothesis is true, when we reject it we commit what is called a type 1 error.

Suppose that the true detection percentage is 70% and we set the rejection region at 18 or more cases in a sample of 20; we can use the equation defining the binomial probability model to show that there is a probability of 0.0355 that we will obtain 18 or more abnormal ultrasounds. If we had set the rejection region at 19 or more cases, then from the equation defining the binomial probability model, we see that there is a probability of 0.0076 of observing data in the rejection region. By picking a rejection region before we carry out the study (e.g., 18 or more abnormal ultrasounds), we control the chance of making a type 1 error. Conventionally, rejection regions are set up to keep the type 1 error probability at 5% or lower. This way, when a null hypothesis is true, we will reject it in no more than 5% of studies. One way to keep the type 1 error rate below 5% is to reject the null hypothesis when a $p$ value is $< 0.05$.

### Outline of Remaining Parts of Chapter

In the remainder of this chapter, we illustrate how to analyze and interpret data arising from common research designs, with commonly encountered research questions. We will start with analysis of binary variables and then show the methods for similar designs analyzing continuous variables. All analyses will use a dataset on cartilage degeneration that has 40 individuals each measured at a baseline (time 1) and a follow-up time point (time 2). There are several baseline patient characteristics that do not change over time:

- Age at enrollment (years)
- Sex (male/female)
- Ethnicity (Caucasian/African-American/other)

The outcome is the extent of cartilage degeneration in the knee joint and it is measured at both the baseline and follow-up time points. It is measured in three ways:

- Continuous: area of degeneration in mm$^2$
- Binary: cartilage degeneration vs. no cartilage degeneration
- Ordinal: degeneration graded as limited (0–2), mild (2–4), moderate (4–6), or severe (>6)
- Nominal: cartilage degeneration level (structural, functional, molecular)

The dataset and R code for conducting these analyses can be found on the companion website in the materials for this chapter.
The statistical methods used in the remainder of this chapter are shown in Table 13.2, where they are categorized by the type of outcome and the type of question being asked. The concepts for each of the examples in Table 13.2 will be discussed as they are encountered and we will also show how to carry out each of these analyses using the R programming language.

### Common Study Designs and Analyses

#### Example 1: One Group with Binary Data

We are interested in the underlying probability of cartilage degeneration, based on a sample of patients who have been classified as having degeneration or not. In this example, we illustrate how to:

- Estimate the proportion of cases with cartilage degeneration
- Construct a 95% confidence interval for this proportion
- Test whether this proportion is equal to 0.5

Let \( n \) be the number of patients in the sample and \( x \) be the number with mild to severe degeneration. We estimate the sample proportion as follows:

\[
\hat{p} = \frac{x}{n}
\]

For a binomial random variable the only unknown is the population proportion \( (p) \), and

| Single group | (Example 1) What is the proportion of patients with cartilage degeneration and is it larger than 0.5? | Single-sample binomial test | (Example 5) Estimation of the mean cartilage degeneration level (mm²) along with its 95% CI and a test of whether the mean level of degeneration equals 4 mm² | Single sample t-test. |
| Two groups | (Example 2) Does the proportion with cartilage degeneration differ between males and females? | Two-sample test of proportions; Fisher’s exact test and Pearson’s chi-squared test; odds ratio, risk ratio, and risk difference | (Example 6) Comparison of the mean level of cartilage degeneration level (mm²) between males and females | Unpaired t-test and Mann-Whitney test |
| Multiple groups | (Example 3) Is the proportion with cartilage degeneration different across three age groups? | Pearson’s chi-squared test and Fisher’s exact test | (Example 7) Assess whether the mean level of cartilage degeneration varies across ethnic groups (white, black, and other) | Analysis of variance (ANOVA) and Kruskal-Wallis test |
| One group, two occasions | (Example 4) Is the proportion with cartilage degeneration the same at time 1 as it is at time 2? | McNemar’s test | (Example 8) Assess whether the mean level of cartilage degeneration is the same at time 1 as it is at time 2 | Paired t-test, Wilcoxon sign rank test, and the sign test |

*The corresponding examples are labeled Example 1 to Example 8.*
the standard error of the estimated proportion can be estimated by

\[
SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

When \( n \) is large enough and \( p \) is not close to 0 or 1, \( \hat{p} \) has an approximate standard normal distribution, and hence normal-based methods can be used for hypothesis testing and constructing confidence intervals. In the equation for the 95% CI given below, 1.96 denotes a threshold for a standard normal quantile. In particular, when using a limiting normal approximation, 95% of observations are expected to fall between the ±1.96 limits. Alternative quantiles could be employed, which would yield altogether different confidence coefficients for a specific interval estimation procedure. The 95% confidence interval can be constructed as

\[
\hat{p} \pm 1.96 \times SE(\hat{p})
\]

Suppose we had observed 40 of 40 cases with cartilage degeneration, so that the estimated proportion is 1. If we use the formula above to calculate a 95% confidence interval, we find that it has length zero (i.e., an interval of 1 to 1), because when \( p = 1, p \times (1 - p) = 0 \). This is an example where the standard normal-based approximate confidence interval gives a nonsensical answer; values of the proportion other than 100% could have given rise to the observed value of 100%. One solution to this problem is to add a correction factor of 0.5 to the 40 cases with cartilage degeneration and also the zero cases without. This gives an estimated proportion of \( 40.5/(40.5 + 0.5) = 0.988 \), which can then be used in the standard formula instead of the value of 1, but still using the sample size of \( n = 40 \). The sole advantage of this approach is that it can easily be done with a calculator. A better approach is to use a method of creating the confidence interval that works well in the presence of estimates of 100% or 0%. The R function we use below implements one such method.

For hypothesis testing, our goal is to test:

\[
H_0: p = 0.5 \\
H_a: p \neq 0.5
\]

In R, we need to calculate \( X \) and \( n \) and then we can use a built-in function to estimate \( p \), carry out the hypothesis test and create a confidence interval. In the cartilage dataset, the variable \text{degen} \ has the value “yes” when there is cartilage degeneration and “no” when there is not, so we add up the number of values where \text{degen} \ is “yes” for all subjects to get \( X \). The R code and its output are shown below.

```r
> X <- sum(cart$degen == 'Yes')
> n <- nrow(data)
> binom.test(X,n)

Exact binomial test
data: 28 and 40
corrected number of successes = 27, number of trials = 40, p-value = 0.01659
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval: 0.5346837 0.8343728
sample estimates:
probability of success 0.7
```
Several epidemiological measures of association can be derived from the above table. One measure is the risk difference. We can calculate the risk difference as the proportion of females experiencing cartilage degeneration minus the proportion of males experiencing cartilage degeneration. It is an absolute measure of epidemiological association. Using the notation in Table 13.3, the risk difference (RD) estimator is:

\[ \hat{RD} = \frac{A}{A + B} - \frac{C}{C + D} \]

We estimate each of our independent proportions as outlined in Example 2 and estimate the difference in proportions:

\[ \hat{RD} = \hat{p}_1 - \hat{p}_2 \]

The standard error of RD is given as follows:

\[ \sigma_{\hat{RD}} = \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \]

Assuming that RD has a normal distribution, we can construct a 95% confidence interval:

\[ \hat{RD} \pm 1.96 \sigma_{\hat{RD}} \]

This approach works well with large samples. Other methods may be more appropriate in smaller samples.\(^7\)\(^-\)\(^9\)

Two other measures of association in 2-by-2 tables are the odds ratio (OR) and the risk ratio (RR). In this example, the OR is the ratio of the odds of cartilage degeneration in females to the odds of cartilage degeneration in males. Similarly, the RR is the ratio of the risk (or probability) of developing cartilage degeneration in females to the risk of developing cartilage degeneration in males.

The odds of cartilage degeneration is estimated as \[ \frac{A}{A + B} = \frac{A}{B} \] in females and as \[ \frac{C}{C + D} = \frac{C}{D} \] in males. The odds ratio is:

\[ \hat{OR} = \frac{AD}{BC} \]
233 Statistical Inference: Point Estimation, Confidence Intervals, and Hypothesis Testing

\[ \hat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{AC + AD}{AC + BC} \]

For both the OR and RR, a value of 1 indicates that the likelihood of cartilage degeneration is equal in males and females. A value larger than 1 indicates that the likelihood of cartilage degeneration is larger in females than in males; a value less than 1 indicates that the likelihood of cartilage degeneration is larger in males than in females.

Notice that if the AC term is near zero then the formula for the RR is close to the formula for the OR.\(^\text{10}\) The AC term in the RR formula will approach zero as the number of cases with the disease approaches zero. That is, \(RR = OR\) when the probability of the outcome is low (e.g., when disease prevalence is low). When \(RR > 1\), the OR is larger than the RR; when \(RR < 1\), the OR is smaller than the RR.

It is possible to construct confidence intervals for the RR and OR by hand through use of formulas,\(^\text{7,9}\) but we recommend using built-in functions in R.

To estimate the risk difference we can use the function \texttt{prop.test()} in R, and to estimate odds ratios and risk ratios we can use the \texttt{R} functions \texttt{riskratio()} and \texttt{oddsratio()}.

\[ \text{prop.test(tab)} \]

2-sample test for equality of proportions with continuity correction

data: tab

X-squared = 0.58201, df = 1, p-value = 0.4455

alternative hypothesis: two.sided

95 percent confidence interval:

-0.4974003 0.1741679

sample estimates:

prop 1  prop 2

0.6111111  0.7727273

An alternative to the Pearson's chi-squared test is Fisher's exact test. Note that the odds ratio estimated by the \texttt{fisher.test} function is not the odds ratio we want.
Research Methods in Radiology

> fisher.test(tab)

Fisher’s Exact Test for Count Data

data: tab
p-value = 0.3154
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.0917836 2.2376883
sample estimates:
odds ratio
0.4714461

Finally, to obtain the relative risk and odds ratios comparing women to men on the risk of cartilage degeneration, we can load the epitools package and use the riskratio and oddsratio functions to obtain estimated and 95% confidence intervals (CIs). Both measures indicate a reduction of risk in females; however, the magnitude of the epidemiological measure of association is larger when estimated using the odds ratio estimator compared to the risk ratio estimator. This points out the need to exercise caution interpreting odds ratios. In particular, odds ratios should not be interpreted as risk ratios, unless the underlying probability of event occurrence is small; at which point, estimated odds ratios and risk ratios should be of similar magnitude.

> library(epitools)
> riskratio(tab, conf.level=0.95, rev="both")$measure

risk ratio with 95% C.I.
estimate lower upper
male 1.0000000 NA NA
female 0.7908497 0.5131055 1.218937

> Oddsratio (tab, conf.level=0.95, rev="both")$measure

odds ratio with 95% C.I.
estimate lower upper
male 1.0000000 NA NA
female 0.4748439 0.1100124 1.908347

Example 3: More Than Two Groups with Binary Data

In this example we investigate whether there is an association between cartilage degeneration at time 1 and which of the three age groups the subject is in. Tabulation is an important technique for investigating the relationship between two categorical variables. When we use a hypothesis test to assess whether there is an association between cartilage degeneration and age group, we can think of the same null hypotheses
two ways: (1) \( H_0 \): the proportion with cartilage degeneration is the same in all three age groups; (2) \( H_a \): the proportions in each age group are the same in those with and those without cartilage degeneration. First, we can make and examine a table of degeneration by age group.

```r
> tab <- table(cart$degen1,cart$AgeGroup)
> tab
   <50 50-60 60+
Yes  6  13  9
No   5  4   3
```

To obtain the proportions with degeneration in each age group:

```r
> prop.table(tab, margin=2)
> prop.table(tab, margin=2)
   <50    50-60   60+
Yes 0.5454545 0.7647059 0.7500000
No  0.4545455 0.2352941 0.2500000
```

We can test the hypotheses above using Pearson’s chi-squared test or, when the counts are small, Fisher’s exact test. In R, the chi-squared test function will generate a warning that the test may be inappropriate in small sample size situations.

```r
> chisq.test(tab)
Pearson’s Chi-squared test

data: tab
X-squared = 1.7329, df = 2, p-value = 0.4204
Warning message:
In chisq.test(tab) : Chi-squared approximation may be incorrect
> fisher.test(tab)
Fisher’s Exact Test for Count Data

data: tab
p-value = 0.4667
alternative hypothesis: two.sided
```

With both approaches, we obtain \( p \) values much larger than 0.05, so we do not reject the hypothesis that the different age groups have equal proportions with cartilage degeneration (or that the age group distribution is the same for those with and without cartilage degeneration).

Example 4: Paired Observations with Binary Data

Certain study designs and sampling processes result in dependent data; for example, paired designs where the same individual is measured on separate occasions produce
dependent data. Longitudinal designs and cluster sampling designs are other examples of sampling processes which result in dependent data. So far in this chapter, we have presented methods that are appropriate for data arising from an independent sampling process. If a sampling process results in dependent data, data analytic methods must account for the dependent nature of the data to ensure accurate inferences.

In this example we analyze paired binary data. Specifically, we test whether the proportion of patients with cartilage degeneration is the same at time 1 as it is at time 2 using McNemar’s test. Table 13.4 illustrates how to present categorical data measured on two occasions, tabulating the value at time 1 against the value at time 2.

The total with degeneration at time 1 is $A+B$ and the total at time 2 is $A+C$. The number with degeneration at both times ($A$) is part of both totals, and the difference between these two totals is $B–C$. Only the subjects who change

Table 13.4  Example of a 2x2 table for binary data measured at two time points

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th></th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td>Totals</td>
<td>A+C</td>
<td>B+D</td>
<td>N=A+B+C+D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

> tab <- table(T1=cart$degen1, T2=cart$degen2)
> tab

McNemar’s Chi-squared test with continuity correction

data: tab
McNemar’s chi-squared = 0.1, df = 1, p-value = 0.7518

The proportion of patients with cartilage degeneration at time 1 is $28/40 = 0.70$ and the proportion at time 2 is $30/40 = 0.75$. We do not reject the null hypothesis of equal proportions at times 1 and 2 ($p = 0.7518$).

Example 5: One Group with Continuous Data

In this example we focus on the $N = 40$ measurements of the area of cartilage degeneration, measured in mm$^2$. We assume that our data are independently sampled from a normal distribution with unknown standard deviation that must be estimated from the data. We will show how to:

- Estimate the mean level of cartilage degeneration
- Construct a 95% confidence interval for mean cartilage degeneration
Test the hypothesis that the mean area of cartilage degeneration is equal to 4 mm².

We can estimate the sample mean and sample standard deviation as follows:

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

\[
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

These are used to derive test statistics and confidence intervals based on the t-distribution. Specifically, the quantity below can be used for testing the hypothesis that the mean is equal to \( \mu \). We will use R to carry out hypothesis tests, but show this formula to illustrate three things: (1) The further the observed mean from the hypothesized value, the larger the value of \( T \); (2) the larger the value of \( n \), the smaller the denominator, so the larger the value of \( T \); and (3) the smaller the between-subject standard deviation, the larger the value of \( T \).

\[
T = \frac{\text{Difference}}{\text{Standard Error}} = \frac{(\bar{X} - \mu)}{s / \sqrt{n}}
\]

Using \( T \) as our test statistic we can investigate the following hypothesis:

\[
H_0: \mu = 4 \text{mm}^2
\]

\[
H_a: \mu \neq 4 \text{mm}^2
\]

A single line of R code both estimates the mean and its 95% CI as well as a \( p \) value. By default the \texttt{t.test()} function is used to test whether a vector of observations is equal to zero or not; the argument \texttt{mu=value} can be used to test whether the vector of observations is equal to some nonzero quantity.

```r
> t.test(cart$cart1, mu=4)

One Sample t-test
data: cart$cart1
t = -10.655, df = 39, p-value = 4.12e-13
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 2.158464 2.746099
sample estimates:
  mean of x
  2.452281
```

The sample mean area of cartilage degeneration is 2.45 mm² (95% CI: 2.15–2.74). The null hypothesis value (4 mm²) is not included in the confidence interval, and the one-sample t-test indicates that we reject \( H_0 \) that the true underlying mean is equal to 4 mm² (\( p < 0.0001 \)). This suggests that the true mean level of cartilage degeneration is less than 4 mm².

The t-test assumes that the underlying data are normally distributed. However, it has been shown that the t-test is quite robust to moderate departures from this assumption. Another assumption of the t-test is that the data have been independently sampled. If the data arise from a more complicated sampling scheme—for example, with multiple observations per subject—then estimated standard errors from this simple t-test may be misleading. Care should be taken to ensure that the assumptions are at least roughly met; otherwise, inferences will be wrong (e.g., confidence intervals will not have nominal coverage probability, and hypothesis tests will not have level \( \alpha \) size).
Example 6: Two Groups with Continuous Data

The goal here is to investigate whether the mean area of cartilage degeneration is equal in males and females. To accomplish this task we will also use a two-sample t-test, which estimates the difference in mean cartilage degeneration between the two groups and, as above for the one-sample t-test, divides it by the standard error of this difference. We do not show the formula for the two-sample t-test, as it has the same general form as the one-sample test.

The two-sample t-test allows us to assess whether the mean level of cartilage degeneration is equal in males and females or, as stated below, whether the difference in the means is zero.

\[ H_0: \mu_{\text{Male}} - \mu_{\text{Female}} = 0 \]
\[ H_a: \mu_{\text{Male}} - \mu_{\text{Female}} \neq 0 \]

In addition to the assumptions specified for the one-sample t-test, the standard two-sample t-test also assumes equal variances in the two groups (i.e., the variance of one group cannot be too different from that of another group). To protect ourselves from errors if this assumption is incorrect, we recommend using the version of the t-test that allows unequal variance (i.e., the Welch two-sample t-test).

```
> t.test(cart1~gender, var.equal=FALSE, data=cart)

Welch Two Sample t-test

data: cart$cart1 by cart$gender
t = -1.9016, df = 35.729, p-value = 0.06531
alternative hypothesis: true difference in means not equal to 0
95 percent confidence interval:
 -1.11572038 0.03606383
sample estimates:
mean in group female mean in group male
2.155376 2.695204
```

The results above show that the males have a slightly larger mean than the females, with a p value of 0.06531 and a 95% confidence interval for the difference running from –1.116 to 0.0361. If the outcome variable is too far from normally distributed, or to check the robustness of the finding from the two-sample t-test, we can also compare a continuous outcome in two groups using what is called a nonparametric approach. The Wilcoxon rank sum test (also known as the Mann-Whitney test) is not based on the observed areas of degeneration, but on their ranks—the smallest value is given a rank of 1, the second smallest a rank of 2, and so on. This approach does not test the equality of means, but rather the equality of the distributions, for example, whether one group’s distribution is shifted up or down relative to the other. The Wilcoxon rank sum test is implemented using the following R code, and gives a conclusion similar to the t-test, with only a small change in the p value and 95% CI.

```
> wilcox.test(cart1~gender, data = cart, conf.int = TRUE)

Wilcoxon rank sum test

data: cart1 by gender
W = 132, p-value = 0.07476
```
Example 7: More Than Two Groups with Continuous Data

Commonly, a researcher wants to assess whether the means or distributions of more than two groups differ. Analysis of variance (ANOVA)11 and its nonparametric equivalent analogue (the Kruskal-Wallis test)12 extend the ideas of t-tests and Wilcoxon rank sum tests to more than two groups. In this example, we investigate whether the mean cartilage degeneration (mm²) is the same in three ethnic groups (Caucasian, African-American, and other) versus the alternative hypothesis that at least one group mean is different from the others. The mathematics of the approach involve intermediate calculations that are typically not carried out directly by the analyst, but within the statistical software being used. For those interested in the details, we refer the reader to Rosner1 and for more applications to R, Dalgaard.2

Formally our null and alternative hypotheses for the means in the three groups are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 \]
\[ H_a: \text{At least one of the means } \mu_j \text{ is not equal to the others} \]

In R we can plot the distribution of cartilage degeneration in each ethnic group using side-by-side box plots. This graphical summary suggests that cartilage degeneration is highest in Caucasian individuals. We can make the boxplots in Fig. 13.4 and test the hypothesis of equal means in all three groups using the following R code:

```
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 
  -1.05546453  0.06785573
sample estimates:
difference in location 
  -0.4802262
```

![Fig. 13.4](image-url) Side-by-side box plots showing the distribution of the area of cartilage degeneration in each ethnic group.
To find out which groups are statistically significantly different, we can make all three pairwise comparisons between groups using the `pairwise.t.test` command:

```r
> pairwise.t.test(cart$cart1,cart$ethnicity)

Pairwise comparisons using t tests with pooled SD
data: cart$cart1 and cart$ethnicity

African-American   Caucasian
Caucasian           0.095           -
Other              0.586           0.031

P value adjustment method: holm
```

This suggests that at least one of the means is different from another—observed differences this large (or larger) between the groups would happen only with a probability of 0.0263 if the means were all the same.

To find out which groups are statistically significantly different, we can make all three pairwise comparisons between groups using the `pairwise.t.test` command:

```r
> pairwise.t.test(cart$cart1,cart$ethnicity)

Pairwise comparisons using t tests with pooled SD
data: cart$cart1 and cart$ethnicity

African-American   Caucasian
Caucasian           0.095           -
Other              0.586           0.031

P value adjustment method: holm
```

After accounting for multiple comparisons using the Holm method, we find that only the Caucasian and the Other ethnic groups are statistically different with respect to the extent of cartilage degeneration.

ANOVA makes many of the same assumptions as the two-sample t-test: the outcomes are normally distributed; the variances of the outcomes within each group are equal, and the data are independent samples. As with the t-test, with larger sample sizes in each group, ANOVA becomes less sensitive to violations of these assumptions.

However, if the normality assumption is strongly violated, the nonparametric Kruskal-Wallis test can be used instead of ANOVA. As with the Kruskal-Wallis test, it replaces the observed cartilage degeneration values by their ranks in the sample and it does not test equality of means, but equality of distributions across the three groups.

To run the Kruskal-Wallis test in R and follow it up with pairwise Wilcoxon rank-sum tests, run the R code below. The results are very similar to those of the ANOVA procedure.

```r
> kruskal.test(cart1~ethnicity, data=cart)

Kruskal-Wallis rank sum test
data: cart1 by ethnicity

Kruskal-Wallis chi-squared = 5.6896, df = 2, p-value = 0.05814
```
This example demonstrates that the results of hypothesis testing can depend on the methods used. With ANOVA, we reject the hypothesis that mean cartilage degeneration is equal across ethnic groups and through estimation of pairwise differences find that the “other” and Caucasian groups are different, all at the $\alpha = 0.05$ level. By contrast, with the same $\alpha$, the Kruskal-Wallis method does not reject the null hypothesis of equal distributions of cartilage degeneration across the three ethnic groups ($p = 0.05814$). With that finding, there is not a strong justification for making pairwise comparisons across the three ethnic groups. If do go ahead and make those comparisons, we find that none of the pairwise comparisons has $p < 0.05$. It may be tempting, when there are multiple methods for carrying out the same kind of analysis, to investigate results under each method and pick the one yielding the most pleasing results, usually the ones with the smallest $p$ values. This is an example of poor research/ethical practice. One should not choose methods according to the statistical significance of their findings. Rather, we recommend that when different approaches give different results, that the investigator tempers his/her enthusiasm for the most positive findings and examines a number of things. For this example: (1) Are the assumptions of the ANOVA met, such that the variability is the same across groups? Fig. 13.4 suggests that there is most variability in the “Other” group and least in the African-American group. (2) Are there one or more large or small values that unduly affect the means in the ANOVA? In the nonparametric analysis based on ranks, these values are not as influential. (3) Is it simply that the ANOVA and Kruskal-Wallis test are answering slightly different questions, one comparing means and the other comparing distributions?

**Example 8: Paired Observations with Continuous Data**

In this example, we introduce three methods for conducting analyses related to paired means (i.e., means of the same variable measured on the same subject at two different time points or by two different assessors): the paired t-test, Wilcoxon’s signed rank test, and the sign test. We will apply these data to the cartilage degeneration data obtained on subjects at time points 1 and 2 and assess whether the mean cartilage degeneration is the same at time 1 and time 2.

The t-test for paired data proceeds in two steps: first, find the within-subject difference (the time 2 value minus the time 1 value); use a single sample t-test to test the hypothesis that these differences have a true mean of zero. That is, we can define a new parameter $\mu_D = \mu_2 - \mu_1$, and then test

$$H_0: \mu_D = 0$$

$$H_a: \mu_D \neq 0$$

The t-statistic is the same as the one for the single-sample t-test in example 5, except that the observations are the within-subject differences.

We can carry out a paired t-test in R using either of the two lines of the following code, taking care to use the paired=TRUE argument in the first option.
The results show that the mean area of cartilage degeneration is −1.18 mm² lower at time 1 compared to time 2 (95% CI: −1.67, −0.69). The null hypothesis value of zero is not included in the 95% confidence interval, a finding that is concordant with the small p value from the hypothesis test. Therefore, at an α level of 0.05, we reject the null hypothesis and conclude that the mean level of cartilage degeneration at time 2 is higher than that at time 1 (p < 0.0001).

As with the t-test for the comparison of two groups, the paired t-test makes certain assumptions about the distribution of the within-subject differences; most importantly the results will be valid when these differences have a close to normal distribution. We present two alternative methods which may be used to compare paired continuous data: the Wilcoxon signed rank test and the sign test. These can be implemented in R as follows, and, not surprisingly, given the very strong evidence of a difference from the paired t-test, they give results that support the inferences generated by the paired t-test.
The sign test works from the assumption that when the distributions of cartilage degeneration area at time 1 and time 2 are the same, 50% of subjects will have larger time 1 values and 50% will have larger time 2 values. We count the number of observations where time 2 is larger and then use a single sample test of whether the true proportion is 0.5. The R code below runs this analysis and finds that 27 of 40 (67.5%) subjects had larger time 2 values; the $p$ value for this, under the hypothesis that the true proportion is 50%, is 0.03848.

```r
> timeTwoBigger <- sum(cart$cart1<=cart$cart2)
> n <- length(cart$cart2)
> binom.test(timeTwoBigger, n)

Exact binomial test

data: timeTwoBigger and n
number of successes = 27, number of trials = 40, p-value = 0.03848
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5087051 0.8142710
sample estimates:
probability of success
 0.675
```

### Highlights of Key Points

- **Point estimation** concerns using information from a sample of data to estimate a fixed/unknown (population) parameter of a statistical model (e.g., the mean of a normal model, or the proportion in a Bernoulli/Binomial model).
- **Interval estimation** finds a plausible range of values—the confidence interval—for the true value of a parameter.
- **Hypothesis testing** is concerned with deciding whether an underlying true value is different from a hypothesized value.
- The study design and type of outcome variable influence the choice of the most appropriate methods for estimation and for hypothesis testing. This chapter has presented widely used methods for binary and continuous outcomes for a single sample, two samples, multiple samples, and paired another designs.

### References

Learning Objectives

- Introduce simple linear regression, multiple linear regression, and logistic regression.
- Discuss the parameterization of each of the above models and interpretation of estimated regression coefficients.
- Introduce Pearson’s and Spearman’s correlation.

Introduction

The purpose of this chapter is to introduce readers to commonly encountered regression models in biomedical research. In discussing regression, the relationship between the simple linear regression model and Pearson’s measure of correlation is touched upon. We briefly introduce the nonparametric Spearman’s correlation coefficient. Finally, the logistic regression model is discussed as it is used to assess covariate effects on a binary outcome variable. We demonstrate the computational implementation of these statistical models using R. We provide illustrative examples of each of these methods using the cartilage degeneration data set introduced in previous chapters.

Simple Linear Regression

Simple linear regression seeks to describe the dependence between a single outcome variable (y) and a single covariate (x). The outcome variable (y) is sometimes also referred to as the response or dependent variable. The covariate (x) is sometimes referred to as the independent, predictor, or explanatory variable. For example, if we consider the cartilage degeneration data set, that was introduced in Chapter 13, simple linear regression could be used to investigate the association between extent of cartilage degeneration (our outcome variable) and age (our covariate). In particular, we are interested in assessing the extent to which the mean level of cartilage degeneration changes as we age. In this particular example, both the outcome variable and the covariate are measured on a continuous scale. However, covariates may be of any type when handled appropriately, as discussed later in the chapter. When the outcome variable is not continuous (or not normally distributed), then linear regression is not an appropriate regression model. Depending on the measurement scale of the outcome variable, an extended class of regression model, that is, generalized linear models (GLMs), may be more appropriate. A thorough treatment of GLMs is given in McCullagh and Nelder. We explore one particular class of GLM suitable for binary outcome data: the logistic regression model. We do not cover any other GLM classes that may have important applications for modelling count based responses, categorical responses, etc.

The primary reason for fitting a simple linear regression model to data is to explore/estimate the linear dependence between covariate and outcome. Considering the cartilage degeneration example, we may use linear regression to investigate the following research question: Is there a linear relationship between patient’s age (predictor variable) and area of degeneration (outcome variable) as in Fig. 14.1?

The simple linear regression model can be represented in a manner such that a particular value of slope and a particular value of intercept defines the line (the overall linear trend) that passes as close as possible to all of
the points of the scatterplot shown in the left panel of Fig. 14.2.

```
plot(x=X$age,y=X$cart2,xlab="Age (years)", ylab="Cartilage Degeneration (squared mm)"
```

Obviously, many possible straight lines will fit the observed data reasonably well (and even more could fit the data poorly). The goal will be to find the best fitting line for these observed data. If every data point fit on a line, then every data point \((x_i, y_i)\), would satisfy the equation \(y = \beta_0 + \beta_1 x\) for some appropriate value of the intercept, \(\beta_0\), and some appropriate value of the slope, \(\beta_1\). It is rare that observed data points fall exactly along the fitted regression line (see Fig. 14.2). The vertical
distance between observed data points and the model fitted value is called the residual/error (and is denoted $e_i$). As such the linear model can be defined as:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where $y_i$ represents the outcome variable (for subject $i = 1,..., n$), and $x_i$ represents the covariate value (again, for subject $i = 1,..., n$). The $e_i$ term represents an error term (or residual) (for subjects $i = 1,..., n$), and the parameters $\beta_0$ and $\beta_1$ are the intercept and slope, respectively. Isolating for the error term yields:

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

Therefore, the error term is the difference between the actual observed value of the outcome ($y_i$) and the corresponding value on the line ($\beta_0 + \beta_1 x_i$) as illustrated in the right panel of Fig. 14.2. In the simple linear regression model, parameters are chosen such that they minimize the sum of these squared values between the observed data and the fitted model values (i.e., they minimize the sum of the squared residuals).

### Interpreting the Model

In R, linear model estimates are easily obtained with the `lm` command.

```r
> summary(lm(X$cart2 ~ X$age))
```

Call:
`lm(formula = X$cart2 ~ X$age)`

Residuals:
Min 1Q Median 3Q Max
-1.9924 -0.8880 -0.1599 0.7259 4.1834

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.341 on 38 degrees of freedom
Multiple R-squared: 0.4539, Adjusted R-squared: 0.4396
F-statistic: 31.59 on 1 and 38 DF, p-value: 1.889e-06

The least squares estimates corresponding to the slope and intercept parameters are 0.15 and -4.6, respectively. The units of the intercept are the same as the unit of the outcome, which is mm$^2$ in this case. The units of the slope, however, is a ratio of the outcome’s units and the covariate’s units. In this case, it would be mm$^2$/years. The slope parameter from the simple linear regression model fit denotes the expected change in the mean of the outcome variable for a unit increase in the covariate. In our example, a year increase in age is estimated to increase the mean level of cartilage degeneration by 0.15 mm$^2$. The intercept in a simple linear regression model can be interpreted as the expected mean of
the outcome when the covariate is held at zero. Depending on the measurement scale of the covariate, this may or may not make sense for a particular application. In our example we have said that the mean level of cartilage degeneration for someone of age zero is −4.59 mm². This illustrates one danger when extrapolating the model, or making conclusions beyond the scope of the data. The age of patients in the data does not extend to 0 years so extending the line that far to make conclusions is dangerous. One common way to give meaning to the intercept is to shift the model around the mean values of the predictor variable by subtracting the predictor’s mean from each predictor value: \( x_{i}^{(\text{new})} = x_i - \bar{x} \) and using the values of \( x_{i}^{(\text{new})} \) rather than of \( x_i \) in the model. The slope will remain unchanged by this transformation of the covariate. The intercept is still measuring the mean value of the outcome when the covariate is held at zero; however, under this covariate transformation, zero represents the mean value of age. So, in this example, the intercept now represents the average outcome for an individual whose age is equal to the mean empirical age observed in the sample.

**■ Inference**

The previous sections introduced the simple linear regression model, discussed parameter estimation and model interpretation. Here we consider two aspects of inference for simple linear regression: confidence interval estimation and hypothesis testing.

In the context of the linear regression model, we are interested in performing inference on the intercept and slope parameters, respectively. In particular, we are primarily interested in testing whether the values of a parameter equals some specific quantity. For example, focusing on the slope parameter one is typically interesting in testing:

\[
\begin{align*}
H_0: & \quad \beta_1 = 0 \\
H_1: & \quad \beta_1 \neq 0
\end{align*}
\]

Normal sampling theory can be used to demonstrate that the distribution of the associated test statistic under the null hypothesis follows an exact t-distribution (with \( n-2 \) degrees of freedom, as the regression model includes two parameter estimates). Our t-test statistic is given below:

\[
T_{n-2} = \frac{\hat{\beta}_1 - \beta_0}{se_{\hat{\beta}_1}}
\]

T-based confidence intervals can be obtained inverting the aforementioned test statistic, yielding a t-based interval:

\[
\hat{\beta}_1 \pm t_{(n-1, \frac{1}{2})} \times se_{\hat{\beta}_1}
\]

Based on the output from the simple linear model, \( \hat{\beta}_1 = 0.15054 \) and \( se_{\hat{\beta}_1} = 0.02678 \). The critical value for a 95% confidence interval on 38 degrees of freedom (because \( n = 40 \)) is \( t = 2.024 \). Using the estimated slope parameter, its corresponding standard error and noting the critical value of the \( t_{38} \) distribution, we can go about constructing confidence intervals and performing hypothesis testing. Using the t-based approach defined above, we estimate a 95% confidence interval for the slope parameter to be (0.096, 0.205). The corresponding test statistic is 5.62, with an associated p-value of \( 1.86 \times 10^{-6} \). The p-value is less than the traditional \( \alpha = 0.05 \) level for declaring statistical significance. We note that the confidence interval does not include zero. These two facts suggest that we should reject the null hypothesis that \( \beta_1 \) is zero. We conclude that a unit increase in age results in a statistically significant, 0.15 unit increase in expected cartilage degeneration.

**■ Pearson’s and Spearman’s Correlation Coefficients**

A quantity describing the extent of linear dependence between two variables is Pearson’s correlation coefficient. Pearson’s correlation coefficient assumes that bivariate data exist for a set of \( n \) individuals, that is, we observe pairs of measurements \( (x_i, y_i) \), ..., \( (x_n, y_n) \) on a sample of \( n \) subjects/items. The linear correlation between \( x \) and \( y \) as measured...
by Pearson's correlation coefficient is expressed as:

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]

The correlation coefficient is a scale-free number ranging between -1 and 1. When the correlation coefficient is negative, it suggests that if one variable increases then the other variable will decrease. If the correlation coefficient is positive, then if one variable increases then the other will increase as well.

The following equation illustrates the relationship between estimated model parameters, \( \beta_0 \) and \( \beta_1 \), and other sample quantities such as: the standard deviation of the outcome (SDY), the standard deviation of the covariate (SDX) and the correlation coefficient \( (r_{xy}) \). It also reveals an interesting relationship between linear regression model estimators and Pearson's correlation coefficient \( (r_{xy}) \):

\[ \hat{\beta}_1 = \frac{SD_Y}{SD_X} \times r_{xy} \]

The slope is simply a multiple of the correlation coefficient. Because both standard deviations are positive, the sign of the slope (whether it is positive or negative) matches that of the Pearson's correlation coefficient. In the example, the estimated Pearson correlation coefficient describing the magnitude dependence between age and cartilage degeneration is 0.6737419, indicating a fairly strong positive correlation between the two variables. Pearson's correlation coefficient can be estimated in R using the `cor(..., method="pearson")` function.

A criticism of the Pearson's correlation coefficient is that it is sensitive to outlying values. An alternative approach to measuring the dependence between two variables is the Spearman's rank correlation coefficient, denoted by \( r_s \). The Spearman's coefficient models the degree of a monotonic relationship observed between \( x \) and \( y \). Spearman's correlation is calculated using the same equation as was introduced for Pearson's correlation but applied to the ranks or order statistics, rather than to the values of the data themselves. This means that Spearman's rank correlation can be applied on variables measured on an ordinal scale.

For this example, the estimated Spearman correlation coefficient between age and cartilage degeneration is 0.6525882. Spearman's correlation coefficient can be estimated in R using the `cor(..., method="spearman")` function. A summary of approaches that can be used to assess relationships between variables is shown in Fig. 14.3.

**Multiple Linear Regression**

Multiple linear regression extends the simple linear regression model to instances where there are two or more covariates. Below is an example where the multiple linear regression model includes \( p \) covariates and an intercept, leading to \( (p + 1) \) regression coefficients. In this case, for every observation, there must now be a value measured for the outcome variable, and a value measured for each of the \( p \) covariates included in the model. The form of the multiple linear regression model is

\[ y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \]

where the different \( \beta_i \) regression parameters are slopes corresponding to a given variable \( x \), for \( j = 1, 2, \ldots, p \). \( \beta_0 \) is the intercept and \( \epsilon \) is the error term. The construction and interpretation of error terms, hypothesis tests, confidence intervals, etc. remain largely unchanged from what was discussed in the simple linear regression setting. As the multiple linear regression model consists of \( p > 1 \) slope parameters, we introduce the \( F \)-test as a method for assessing whether multiple regression parameters are simultaneously zero. Mathematically our null hypothesis is:

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \]

\[ H_1: \text{At least one } \beta_j \text{ is non-zero} \]

This is sometimes referred to as a global \( F \)-test. \( F \)-tests corresponding to a hypothesis that a subset of parameters in the regression model equals zero are also possible. The idea of the \( F \)-test is to fit two models to the data: the full
model with $p$ covariate parameters, and the restricted model with $(p - k)$ parameters, which is a model where $k$ of the predictors are assumed to have no effect, so their regression coefficients are equal to zero. Each model will have a residual sum of squares (RSS) value associated with it, reflecting the degree to which the proposed model fits the data. If the $k$ proposed constraints on the model have no effect on fit (the omitted predictors did not linearly explain much variability in the outcome), then one expects to observe similar RSS values between the two model fits. However, if the additional $k$ terms in the full model do improve fit substantially, then this would be reflected by a smaller RSS value. The null hypothesis (that the $k$ slopes are equal to some constant values, usually all 0) is rejected if one observes a large $F$-statistic (i.e., the observed $F$-value exceeds the critical threshold defined as the value of the $F$-test under the null hypothesis).

Returning to the cartilage degeneration example, we regress cartilage degeneration on age (continuous), as well as gender (nominal) and ethnicity (nominal) (Fig. 14.4). Gender is a two-level categorical variable and ethnicity is measured as a three-level categorical variable. Categorical variables are handled slightly differently than continuous variables when they are included as covariates in a regression model.

A common approach to representing categorical variables in a regression model is referred to as dummy variable coding. If a categorical variable has $k$ levels then one introduces $k - 1$ dummy variables into the model. Gender has two levels. Therefore, one would need a single dummy variable that is equal to 1 if the observation corresponds to a preselected gender, and 0 for the other gender. One can choose either the male or female level of the gender variable to act as the reference/baseline gender. Suppose the male category is chosen as the baseline/reference value, then the dummy variable equals 0 for observations corresponding to males and 1 for observations corresponding to females. Although estimates from the model will differ depending on which categorical level was considered referent category, the final conclusions and interpretations will always be the same, so the choice of referent category is arbitrary. Similarly, for a three-level categorical variable like ethnicity, dummy variables are introduced to represent when the observations correspond to Caucasians, African-Americans, or others. Supposing that Caucasian is chosen as the baseline for ethnicity, one dummy variable is created that is equal to 1 if the observation refers
to an African-American, and is equal to 0 otherwise, and another dummy variable is created that is equal to 1 if the observation refers to an ethnicity that is listed as "other" and is equal to 0 otherwise. The two dummy variables clearly can never simultaneously be 1, but can simultaneously be 0, indicating the observation referred to a Caucasian. These two dummy variables are included in the regression equation. After regressing cartilage degeneration on gender (one dummy variable for gender) and ethnicity (two dummy variables for ethnicity), the parameter estimates from the resulting R model fit are given below. Note that the regression model has four covariates in total.

```r
> summary(lm(X$cart2 ~ X$age + I(X$gender=="female") + I(X$ethnic=="african-american") + I(X$ethnic=="other")))
```

**Coefficients:**

| Predictor variable | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------|----------|------------|---------|----------|
| (Intercept)        | 5.538722 | 1.719911   | 3.220   | 0.00276  ** |
| X$age              | -0.004983| 0.029678   | -0.168  | 0.86764  |
| I(X$gender == "female")TRUE | -3.199305 | 0.484953 | -6.597  | 1.27e-07 *** |
| I(X$ethnic == "african-american")TRUE | -0.293739 | 0.364579 | -0.806  | 0.42586  |
| I(X$ethnic == "other")TRUE | -0.293380 | 0.342827 | -0.856  | 0.39795  |

**Residuals:**
Min 1Q Median 3Q Max
-1.5242 -0.5286 0.0241 0.3944 2.2334

Fig. 14.4 Overview of variables in a multiple linear regression model.
Research Methods in Radiology

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8033 on 35 degrees of freedom
Multiple R-squared: 0.8196, Adjusted R-squared: 0.799

F-statistic: 39.76 on 4 and 35 DF, p-value: 1.478e-12

The F-test from the model is used to test the global null hypothesis that all of the regression coefficients are equal to 0 against the alternative hypothesis that at least one of the regression coefficients is nonzero. The p value from the F-test is $1.478 \times 10^{-12}$ suggesting that we reject the null hypothesis at the usual 5% level. Proceeding, one can begin to interpret each of the estimated regression coefficients, their standard errors, test-statistics, and corresponding p values. The estimated intercept is 5.538722 (SE = 1.719911). The p value associated with the intercept is 0.00276; therefore, at a 5% alpha level, we reject the null hypothesis that the intercept is 0. The interpretation of the intercept corresponds to the mean level of cartilage degeneration when all variables are 0 (age = 0, gender = male, ethnicity = Caucasian). The adjusted change in cartilage degeneration expected for each 1-unit increase in age is $-0.004983$ ($p = 0.86764$). For the categorical variables, the estimated regression coefficients are interpreted in relation to the dummy variable. In other words, the mean level of cartilage degeneration for females is 3.199305 mm$^2$ lower than for males ($p = 1.27 \times 10^{-7}$), after considering other variables in the regression model. Had the dummy variables been coded the opposite way (the dummy variable = 0 if female and = 1 if male), the slope would be +3.199305 but the interpretation is the same with the same p value. Similarly, the mean level of cartilage degeneration is 0.293739 ($p = 0.42586$) units lower for African-Americans and 0.293380 units lower for persons from other ethnic groups ($p = 0.39795$) compared to those of Caucasian ethnicity—again, after adjusting for other variables in the model.

Regression model diagnostics, which validate that the regression model fits the data well, are an important part of data modeling. There are many different approaches to assessing model goodness of fit, and a thorough treatment is not possible in one chapter. A good reference on regression model diagnostics is Weisberg.2 In R, one can quickly obtain some plots, as shown in Fig. 14.5, which can be used to assess some aspects of goodness of fit. For example, the plot at the top left of the residuals versus fitted values can be used to assess whether there are any model predicted/fitted quantities which are overly large or small. Here, most values are between −2 and +2 which is reasonable. The normal quantile-quantile (QQ) plot can be used to assess if the (standardized) residuals from the model fit are roughly normally distributed. If the (standardized) residuals are roughly normally distributed then the points in the QQ-plot will lie on the line $y = x$. If the (standardized) residuals are not from a normally distributed population, then points in the QQ-plot will exhibit a nonlinear relationship. The bottom left plot is similar to the top left, illustrating the relationship between scaled/standardized residuals and fitted values. The plot on the bottom right is useful for determining if the linear model fit (i.e., the regression coefficients) are being influenced by some small number of outlying or otherwise influential observations (which, in the example, may be the case with subject = 22).

Multiple Logistic Regression

Many outcome variables encountered in applied biomedical research are dichotomous, such as diseased versus not diseased, dead versus alive, etc. Logistic regression provides
a unified framework for assessing the impact of one or more covariates on a dichotomous outcome variable.

To begin, suppose the outcome variable, \( y_i \), takes on two possible values: 0 and 1. Although the outcome is 0 or 1, the probability of it being 1 could be any number between 0 and 1. Letting the probability that \( y_i = 1 \) be \( p_i \),

\[
P(y_i = 1) = E(y_i) = p_i
\]

the expected value of \( y_i \) equals the probability that \( y_i = 1 \) which equals \( p_i \). The probability scale (between 0 and 1) can be transformed to be any number by now considering the odds of \( y_i = 1 \), which is \( \frac{p_i}{1-p_i} \), and then taking the log of this ratio. The result is called the logit which is represented by \( \log \left( \frac{p_i}{1-p_i} \right) \). The logistic regression model links the log odds of \( p_i \) to covariates as follows:

\[
\log \left( \frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}
\]

Isolating for \( p_i \) results in:

\[
p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}
\]
which is called the logistic function. Note that exp() denotes exponentiation with base e. The first equation linearly relates the log of the odds of an event to covariates, and so the interpretation of slopes differs slightly. Exponentiating one of the parameters estimated on the log-odds scale results in a transformed parameter which can be interpreted as an odds ratio. When the estimated parameter $\beta$ is negative, a unit increase in the covariate results in a $100 \times (1 - e^\beta)$ percent decrease in the odds of the outcome (this is true for continuous covariates and for dummy coded categorical covariates coded as 0 or 1). When the estimated parameter $\beta$ is positive, a unit increase in the covariate results in a $100 \times (e^\beta - 1)$ percent increase in the odds of outcome. After obtaining estimates of regression parameters in the multiple logistic regression model, one typically proceeds to conduct inference on estimated quantities via confidence intervals and hypothesis tests. The machinery introduced in multiple linear regression carries over to multiple logistic regression. T-tests and t-based confidence intervals for single regression parameters are replaced with standard normal approximations (Z-tests, etc.). F-tests are replaced with asymptotic chi-squared tests.

Below we illustrate how logistic regression can be used to investigate the impact of age on the probability of developing mild/moderate/severe cartilage degeneration. Note that our outcome is a binary composite: mild/moderate/severe degeneration vs. no degeneration. Because logistic regression, like other types of regression, is a special case of the generalized linear model, it can be fit with the glm() function in R. The glm() function can be used to fit a broad class of generalized linear models, of which logistic regression is a specific type. To fit a logistic regression model in R using the glm() function, the user must specify that the assumed distributional family for the outcome is “binomial” and that the link function is the logit link. Other values for these keyword arguments will result in fitting different classes of GLMs.

```r
> summary(glm(I(X$cart2_nom=="mild/moderate/severe") ~ X$age,
family=binomial(link="logit")))
```

Call:
```r
glm(formula = I(X$cart2_nom == "mild/moderate/severe") ~ X$age,
family = binomial(link = "logit"))
```

Deviance Residuals:
```
Min 1Q Median 3Q Max
-1.67507 -0.05446 0.28454 0.53244 1.84873
```

Coefficients:
```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.96265   3.82016  -2.608  0.00911 **
X$age        0.21238   0.07531   2.820  0.00480 **
```

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 44.987 on 39 degrees of freedom
Residual deviance: 31.819 on 38 degrees of freedom

AIC: 35.819

Number of Fisher Scoring iterations: 5

The regression coefficients from a logistic model fit are interpreted differently from those of the linear model fit discussed above. In logistic regression the raw regression coefficients are measured on a log-odds scale. In other words, the intercept represents the log-odds of having mild/moderate/severe cartilage degeneration

...
include multiple variables. If ethnicity (baseline = Caucasian, and appropriate dummy variables for other) is included in the multiple logistic regression model, one can investigate how age impacts the odds of mild/moderate/severe cartilage degeneration after adjusting for this second variable. The R code and resulting logistic model fit are given below:

```r
> glm.fit2 <- glm(I(X$cart2_nom=="mild/moderate/severe") ~ X$age + I(X$ethnic=="african-american") + I(X$ethnic=="other"), family=binomial(link="logit"))
> exp(cbind(OR = coef(glm.fit2), confint(glm.fit2)))

Waiting for profiling to be done...

OR 2.5 % 97.5 %
(Intercept) 6.235829e-06 3.344362e-11 0.02074097
X$age 1.366152e+00 1.146349e+00 1.83337826
I(X$ethnic == "african-american") TRUE 1.162980e-02 1.149413e-04 0.26340909
I(X$ethnic == "other")TRUE 4.197326e-02 6.555952e-04 0.83081830

Here, a unit increase in age results in a 1.366152 unit increase in the odds of mild/moderate/severe cartilage degeneration after adjusting for ethnicity. For categorical variables the estimated odds ratios are interpreted for each coefficient in relation to the reference level. In this example, African-Americans have approximately 0.0116 times the likelihood of mild/moderate/severe cartilage degeneration and the other ethnic group has 0.0419 times the likelihood of mild/moderate/severe cartilage degeneration, after adjusting for age. The confidence intervals for each of these regression coefficients include 1; therefore, neither term is a statistically significant predictor of the outcome.

A similar approach to linear regression extends the logistic regression model to include multiple variables. If ethnicity (baseline = Caucasian, and appropriate dummy variables for other) is included in the multiple logistic regression model, one can investigate how age impacts the odds of mild/moderate/severe cartilage degeneration after adjusting for this second variable. The R code and resulting logistic model fit are given below:

```r
> glm.fit1 <- glm(I(X$cart2_nom=="mild/moderate/severe") ~ X$age, family=binomial(link="logit"))
> exp(cbind(OR = coef(glm.fit1), confint(glm.fit1)))

Waiting for profiling to be done...

OR 2.5 % 97.5 %
(Intercept) 4.712772e-05 5.471628e-09 0.03111613
X$age 1.236617e+00 1.090338e+00 1.48171859

A similar approach to linear regression extends the logistic regression model to
Conclusion

This chapter introduced readers to regression models including multiple linear and logistic regression, two workhorses in applied biostatistics. Two important areas that received little attention in this chapter were model diagnostic checking and model selection. Both of these topics have received book-length attention.\(^1\)\(^2\)\(^3\)\(^4\)

In just one chapter, a detailed discussion of all pertinent topics in applied linear and logistic regression modelling is infeasible. Not surprisingly, entire books have been written on each model alone. Some good books include those written by Weisberg,\(^2\) Faraway,\(^3\) Faraway,\(^4\) Harrell,\(^5\) Davison,\(^6\) McCullagh,\(^1\) Vittinghoff,\(^7\) and White.\(^8\) Interested readers should consult these references for a more self-contained discussion of these topics.

Highlights of Key Points

- Linear regression allows one to investigate the impact of one or more covariates on a continuous and normally distributed outcome variable.
- Logistic regression allows one to investigate the impact of one/more covariates on a Bernoulli/Binomial distributed outcome variable.
- Multiple regression extends simple regression via the inclusion of additional (more than one) covariates in the linear predictor of the regression model. This is useful if one wants to obtain the adjusted effect of one variable on the outcome after controlling for the impact of another variable.
- Categorical variables can be included in the linear predictor of a regression model using a dummy variable coding strategy. Other coding strategies exist and are discussed in the provided reference materials.
- Pearson’s correlation can be used to assess linear relationship between two continuous measures. Spearman’s correlation can assess a monotonic relationship between continuous measures.

References

3. Faraway J. Linear Models with R. Boca Raton, FL: Chapman & Hall CRC; 2005
4. Faraway J. Extending the Linear Model with R. Boca Raton, FL: Chapman & Hall CRC; 2006
Learning Objectives

The main goal of this chapter is to explain design and analysis issues that can influence sample size estimation, including:

- Descriptive studies (one-sample)
- Comparative studies (two or more samples)
- Continuous/measured outcomes
- Dichotomous outcomes
- Diagnostic tests
- Reliability studies

Concepts

An important aspect of study design and analysis is sample size. An appropriate sample size achieves certain desirable properties for statistical inferences including descriptive (one-sample) and comparative (two-sample) studies. These can either be for the purpose of estimation of unknown parameters or testing. Table 15.1 briefly reports descriptive and comparative quantities of interest which are suitable for sample size calculations. Sample size is influenced by three major factors: (1) design features, (2) analysis features, and (3) logistics as they shape the problem at hand. Analysis is the focus in this chapter, as design has been previously discussed and logistics are realities that should be incorporated into design and, to an extent, analysis. Table 15.2 provides a brief summary of the three factors.

The motivation for sample size calculations is often to achieve one of two major goals: avoid exceeding a maximum acceptable (as scientifically meaningful) width for a confidence interval, or achieve a predefined power for a study. Intuitively, one should expect that larger sample sizes provide more information and, thus, yield better estimates (narrower confidence intervals) leading to correct decisions in hypothesis tests (adequate power). One enhances the accuracy of estimation through confidence intervals, and the other increases the study power. In either case, the study could be descriptive (one sample) or comparative (two or more groups). Each of these features in various combinations is discussed with examples throughout the chapter.

Sample Size Estimation Based on Confidence Intervals

The first sample size calculations described are those relating to confidence intervals. The general form of a confidence interval for the true mean is given by the following formula:

\[ \text{confidence interval} = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \]

where \( \bar{x} \) is the sample mean, \( Z_{\frac{\alpha}{2}} \) is the (positive) critical value from the normal distribution corresponding to the 100 \((1 - \alpha)\)% confidence interval, \( \sigma \) is the population standard deviation, and \( n \) is the size of the sample. From this formula one can tell that the interval is centered around \( \bar{x} \), and extends a length \( Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \) in both directions from the center, giving a width of \( 2Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \). The width decreases with
increasing values of \( n \). Instead of discussing the full width of the interval, it is common to discuss just the half-width of intervals, \( Z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}} \), how far the interval extends from the center, and is called the margin of error. Therefore, the confidence interval can then be written as

\[
\bar{X} \pm (\text{margin of error})
\]

Both the margin of error and the width of the interval decrease proportionally with \( \frac{1}{\sqrt{n}} \) as illustrated in Fig. 15.1.

Note that this interval width decreases toward 0 as the size of the study sample increases. If \( m \) represents the margin of error of the confidence interval, one can very simply rearrange the formula to obtain

\[
m = Z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{for} \quad n = \left( \frac{Z_{\frac{a}{2}} \sigma}{m} \right)^2
\]

Here, the impact of various factors on sample size is explicit. In particular, to achieving a small margin of error, \( m \), requires a large sample size, \( n \). There are two other major components which impact the width of a confidence interval: the total remaining tail area outside of the interval which defines the confidence interval, \( a \), and the standard error \( \frac{\sigma}{\sqrt{n}} \) (when the central limit theorem is used). The standard error itself is obviously impacted by the standard deviation of the population, \( \sigma \), and the sample size, \( n \).

### Table 15.2 Factors influencing sample size

<table>
<thead>
<tr>
<th>Design</th>
<th>Analysis</th>
<th>Logistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cross-sectional/cohort/case-control</td>
<td>• Scale of outcome</td>
<td>• Drop-out</td>
</tr>
<tr>
<td>• Descriptive/analytical</td>
<td>• Effect measure</td>
<td>• Missing data</td>
</tr>
<tr>
<td>• Parallel/cross-over</td>
<td>• Confidence interval</td>
<td>• Funding</td>
</tr>
<tr>
<td>• Repeated measures</td>
<td>• Test of hypothesis</td>
<td>• Nature of disease</td>
</tr>
<tr>
<td>• Bioequivalence</td>
<td>• Directional vs. nondirectional</td>
<td>• Scope of survey</td>
</tr>
<tr>
<td>• Adjustment for confounders</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 15.1](image) Relationship between confidence interval width and sample size for a 95% confidence interval and known unit variance.
Unlike the hypothesis testing approach, the confidence intervals approach is agnostic to hypotheses. It acts as a descriptive quantity with a particular level of precision. As an example, investigators may want to know how large a study’s sample should be so that the sample mean has, on average, less than a 10% chance of being 3 units away from the true population mean. In this case, the half-width of the confidence interval is 3 units, so the full width of the interval is 6 units, and \( \alpha = 0.1 \) for a 90% confidence interval. Missing here is an estimate of the population’s standard deviation. This may be estimated through a pilot study. The larger the estimate of this standard deviation, the larger the sample size required to ensure that there is less than a 10% chance of the sample mean being 3 units away from the true population mean. That is, a larger sample size is required to ensure that the confidence interval’s width does not grow beyond 6 units. Thus, one may think of the confidence interval of the predefined width as now having a smaller value of \( \alpha \), which means that there is an even smaller chance the sample mean strayed from the true mean than what was originally expected. Typically investigators consider conservative approaches for estimation of sample size, preferring to err on the side of using more resources (larger samples) to ensure that certainty and precision of study results are not compromised.

**Sample Size Estimation Based on Hypothesis Tests (One-sided)**

Hypothesis testing can be another focus for estimating an appropriate sample size for a study. Recall that power is the probability of rejecting the null hypothesis when it is indeed false, and so it is desirable for this to be as close to 100% as much as possible. If investigators use the normal distribution and a predefined power that they were interested in achieving for their study, then the formula for sample size is

\[
\begin{align*}
 n &= \frac{\sigma^2}{(\mu_1 - \mu_0)^2} \left( Z_{1-\beta} + Z_{1-\alpha} \right)^2 \\
 &= \frac{\sigma^2}{(\mu_1 - \mu_0)^2} \left( Z_{\beta} + Z_{\alpha} \right)^2
\end{align*}
\]

where \( Z_{1-\beta} \) is the critical value associated with the power \( 1 - \beta \) (\( \beta \) is the probability of making a type 2 error), and \( \mu_1 \) is the “true” mean used to estimate the detectable effect size of interest from the hypothesized mean, \( \mu_1 - \mu_0 \), and other symbols denote the same quantities as in confidence intervals. Here, one sees that the sample size is influenced as before by the variance of the population, \( \sigma^2 \), namely larger variances demand larger sample sizes, but also by the effect size of interest to detect, and the power. The larger the effect size of interest, the easier it should be to detect, so larger effect sizes decrease the required sample size. Finally, larger values of power correspond to larger values (in magnitude) of \( Z_\alpha \), so this extra certainty in testing requires greater sample size. **Fig. 15.2** illustrates the sample size required for a one-tailed test where the null hypothesis is that the true mean, \( \mu_0 \), is equal to the hypothesized mean, \( \mu_0' \): \( \mu = \mu_0' \) when in truth \( \mu = \mu_1 \neq \mu_0 \) for some \( \mu_1 \).

Unlike the one-sided hypothesis test, there is no formula in the form \( n = \ldots \) for sample size for two-sided tests. Instead, numerical approximations calculated by software are used. These will be investigated directly within examples and exercises.

Variants of formulae for estimating appropriate sample size scenarios are discussed in this chapter, including when standard deviation is estimated rather than known for continuous measurements (using the \( t \) distribution rather than a normal), when the outcome is binary, and performing inference on proportions (using the normal distribution suitable approximations when proportions of interest are not near 0 or 1). Other variants of formulae for sample size estimation arise by making adjustments in the estimate for standard error. The number of variants beyond a simple case is too large to cover exhaustively in this chapter, and usually involve calculations that are tedious, although not difficult. For these and other scenarios, existing software is particularly useful. In this chapter some of the formulae and theory behind these concepts are omitted in favor of illustrating solutions provided by R functions.

**Diagnostic Test Performance Sample Size Calculations**

For sample size estimations of the area-under-the-curve (AUC) and reliability of diagnostic
agreement between two raters when rating images dichotomously can be measured by Cohen's kappa coefficient, which achieves a value of 1 if the two raters completely agree and a value of 0 if the raters only agree as much as expected simply due to chance. If there are more than two possible classifications, say raters are rating the severity on something more like a Likert scale, the intraclass correlation coefficient (ICC) is suitable. The ICC is a ratio (proportion) of variances: the variance between images compared to the total variance, and is the same idea as an ANOVA. If this ratio is close to 1, then most of the variation in ratings is because of the variety of images, meaning that very little is due to variation between raters. The ICC, being a proportion, lies between 0 and 1, but has more sophisticated calculations for sample size as this measures more than just a simple frequency. When incorporating power the numerical methods are also difficult. Although exact methods have been found a useful approximation may be more feasible in practice and only requires reading off published tables. For reliability measured with kappa or ICC, software and tables are referred over formulae.

![Figure 15.2: Relationship between sample size and power.](image)
### Examples Estimating Sample Size

This section exclusively covers a variety of sample size calculations in a variety of contexts. As noted, this is not an exhaustive list but covers many of the concepts used in sample size calculations.

#### Example 1: Target Confidence Interval Width for Count Outcome

Consider a study in which investigators aim to determine the number of metastatic lesions that are detected in the liver in patients with advanced pancreatic cancer by contrast-enhanced CT. The standard deviation of number of lesions per patient is 4.6. On the basis of this standard deviation, what sample size is required for the investigators to construct a 90% confidence interval for the true mean with a margin of error no more than 1.5 lesions per patient?

**Solution:** The solution in this case only requires substituting parameters by numbers into the formula:

\[
 n = \left( \frac{z_{0.05} \cdot m}{s} \right)^2 = \left( \frac{1.6449}{4.6} \right)^2 = 1.5^2 = 25.4
\]

The resultant value is not an integer, which a sample size should be. To obtain a confidence interval no wider than three lesions per patient, one rounds this up to 26. Therefore, at least 26 subjects are required in the study in order to obtain an estimate of the mean number of metastatic lesions per patient detectable by contrast-enhanced CT with a 90% confidence interval no wider than three lesions per patient.

#### Example 2: Target Power for Continuous Outcome—One-Tailed Test

Investigators plan to conduct a study to determine the average size of metastatic lesions in the liver in patients with advanced pancreatic cancer detected by contrast-enhanced CT prior to the start of treatment with a new drug, in comparison to a standard drug. The standard drug provides an average decrease of 23.0 mm in the lesion size but in early trials, the new drug decreased the average size of lesions by 25.2 mm. The standard deviation of decrease in lesion size in the population will be assumed to be 12.1 mm. How many patients must be enrolled in the study to enable us to reject the null hypothesis that the two treatments reduce lesions by the same amount, against the alternative that the new drug reduces the size of lesions even more at the 10% significance level while achieving a power of 80%?

**Solution:** The appropriate sample size is found by substituting parameters by numbers into the formula here:

\[
 n = \frac{s^2}{(\mu_1 - \mu_0)^2} \left( \frac{z_{p} + z_{a}}{2} \right)^2
 = \frac{12.1^2}{(25.2 - 23.0)^2} \left( 0.8416 + 1.2816 \right)^2
 = 136.36
\]

As before, this number is rounded up to achieve a power of 80%. Therefore, 137 patients are required.

#### Example 3: Target Power for Continuous Outcome—Comparison of Two Means

Investigators wish to determine if there is a difference in the mean diffusion tensor MR imaging anisotropy values in patients with multiple sclerosis attending a given hospital as compared with values obtained in the brain MRI of control subjects. Investigators believe that a difference of 50 IU between values of controls and subjects is clinically significant, and from the literature, one expects a standard deviation of 20 IU. What sample size (assuming a 1:1 ratio of cases:controls) is required to detect this difference at the 5% level with at least 90% power?

**Solution:** As mentioned earlier, there is no closed-form formula for the sample size for a two-sided test. In this case, the function in R `power.t.test` is used to provide the desired information using the same information one would expect to use as in a the one-tailed version: effect size to detect, standard deviation, desired power, and significance level.

```r
> power.t.test(50, sd = 22, sig.level = .05, power = .9, type = "two.sample", alternative = "two.sided")
```
Two-sample t-test power calculation

\[ n = 50 \]
\[ \text{delta} = 13.09505 \]
\[ \text{sd} = 20 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.9 \]
\[ \text{alternative} = \text{two.sided} \]

NOTE: \( n \) is number in *each* group

As the bold portion of the output indicates, the study would require 50 cases and 50 controls to achieve a power of 90%.

**Example 4: Target Confidence Interval Width for Proportion Outcome**

Investigating if MRI is more accurate than US to discriminate blood (hemarthrosis) from clear fluid in joints of animals with blood-induced arthropathy, an experiment has been designed to determine the nature of joint material by comparison of results of imaging with those of arthrocentesis of the study joint. Using this as a guide, determine the sample size required to estimate the proportion of cases that represent true hemarthrosis with 80% confidence interval and a margin of error of no more than 5%.

**Solution:** The solution to this modifies the general approach by replacing the estimate for standard error to the appropriate one for proportions. That is, \( \sigma^2 = p(1 - p) \), where \( p \) is the true population proportion. If one had an estimate for the true proportion, it would be used to estimate variance. In lieu of one, the maximum possible variance for proportions is used as a conservative estimate, which corresponds to \( \sigma^2 = 0.25 \). Therefore,

\[
\begin{align*}
    n &= \left( \frac{Z_{0.025}}{m} \right)^2 = \sigma^2 \left( \frac{Z_{0.025}}{m} \right)^2 \\
    &= 0.25 \left( \frac{1.2816}{0.05} \right)^2 = 164.3
\end{align*}
\]

As before, this is rounded up to 164 to ensure at least an 80% confidence interval.

**Example 5: Target Confidence Interval Width for Difference in Proportion—Conservative Estimate**

Investigators plan to determine the difference in accuracy of two diagnostic tests with a 95% confidence interval. This is to be estimated with a margin of error of 10%. Find the sample size to achieve this interval with at least 95% confidence when using the conservative estimate of variance for proportions.

**Solution:** The usual sample size formula is used, replacing the variance with the variance for a difference of independent proportions:

\[
\frac{(Z_{0.025})^2}{(p_1(1 - p_1) + p_1(1 - p_1))} = \frac{1.96^2}{0.1} = 192.08
\]

Therefore, a sample size of at least 192 is required in each sample to achieve a 95% confidence interval that is no wider than 20% for the true difference in proportions.

**Example 6: Target Power for Difference in Proportions—Odds Ratio**

One hundred children with tuberous sclerosis had an MRI of abdomen and were administered gadolinium for diagnosis of renal angiomyolipomas (lesions) in 1995 at a given institution. Of those, 40 presented with chronic nephropathy at the time of the examination and 30 developed nephrogenic systemic fibrosis over 20 years following the gadolinium administration. How many subjects should a research study include to be able to demonstrate a significant difference in number of new cases of nephrogenic systemic fibrosis developed by the patients scanned in the given institution in 1995 who had nephropathy at the time of examination as compared with controls who did not have nephropathy at that time at 5% significance level and a power of 90% if the odds ratio for developing nephropathy when gadolinium is administered to a patient with and without chronic nephropathy is 3?

**Solution:** Because the proportion of patients developing nephropathy (\( p_2 = 0.75 \)) and the odds ratio (\( OR = 3 \)) are given, one can find \( p_1 \) and \( \bar{p} \):

\[
p_1 = \frac{OR p_2}{OR p_2 + 1 - p_2} = 0.9
\]

and using this, the average of \( p_1 \) and \( p_2 \) is

\[
\bar{p} = \frac{1}{2} (0.75 + 0.9) = 0.825
\]
Although the following equation for the one-tailed test could be used to approximate the required sample size,

$$ n = \left( \frac{Z_a \sqrt{2p(1-p) + Z^2 p_1(1-p_1) + p_2(1-p_2)}}{p_2 - p_1} \right)^2 $$

this becomes tedious. Fortunately, this is simple to calculate in R using the default function `power.prop.test`.

```r
> power.prop.test(p1=.9,p2=.75,
sig.level=.05,power=.9,
alternative="two.sided")
```

Two-sample comparison of proportions

```
power calculation
n = 132.7557
p1 = 0.9
p2 = 0.75
sig.level = 0.05
power = 0.9
alternative = two.sided
```

Example 7: Target Power for Difference in Proportions

Consider a study to investigate the difference in the proportion of cases with an accurate diagnosis when contrast was used as compared with the proportion of accurate diagnosis cases when contrast was not used in brain CT scans by gathering an appropriate number of records. If the null hypothesis that the difference between the two proportions is zero and the alternative hypothesis is that the difference is not zero, how many of each type of image pairs (one with contrast and one without) of scans are required for the study to have a type 1 error of no more than 20% and a power of at least 80%? Looking into the literature to obtain approximate estimates for the sample size calculation of a new study, the investigators obtained information from a prior study that approximately 45% of cases were considered normal without and with contrast and 50% of cases were considered normal without contrast and abnormal with contrast. In ideal circumstances, these proportions would have been obtained from independent samples, further motivating the need for this new investigation.

**Solution:** Using R, the solution is given by the `power.prop.test` function:

```r
> power.prop.test(p1=184/400,
p2=202/400,sig.level=0.2,power=0.8)
```

Two-sample comparison of proportions

```
power calculation
n = 898.4248
p1 = 0.45
p2 = 0.50
sig.level = 0.2
power = 0.8
alternative = two.sided
```

Note that the sample size is large primarily because the proportions are so very similar (similar to a small effect size). In order to detect a difference between these two similar proportions (a difference of 5%) the study requires gathering over 899 normal and 899 abnormal images, a total of 1798 records.

Example 8: Target Confidence Interval Width for Kappa

Consider an example of a diagnostic test accuracy study in which two radiologists assess the same set of mammogram images to determine whether the characteristics of detected breast lesions are malignant or benign assuming that abnormal mammograms are expected to have only one lesion per mammogram. Results of the interpretation of mammograms are compared with those of biopsy (reference standard) for corresponding patients. In this study one radiologist deemed 30% of the mammograms to have at least one lesion with malignant characteristics in the set of images evaluated. How many images should be assessed for the investigators to estimate the 95% confidence interval for the true kappa statistic.
to be within a margin of error of 10% if a value of kappa near 70% is anticipated? Assume that biopsies will only indicate that 20% of images actually indicate cancer.  

**Solution:** This question hinges on an appropriate formula for the standard error of Cohen’s kappa, which is a rather involved calculation. Fortunately, many statistical packages include this, and R has a function CIBinary in the kappaSize package which works well for this purpose. Anticipating that $k_0 = 0.7$, kappa0 = 0.7 is specified and the lower (kappaL) and upper (kappaU) limits of the interval are 0.6 and 0.8, respectively, to correspond to a margin of error of 0.1. Further, because the proportion of the population with the trait is 20%, the parameter props = 0.2 is used. Finally, because a 95% confidence interval is desired, alpha = 0.05. Results are shown below.

```r
library(kappaSize)
CIBinary(kappa0 = 0.7, kappaL = 0.6, kappaU = 0.8, props = 0.2, alpha = .05)
```

Hence, a minimum of 376 subjects are required for this study of interobserver agreement.

**Example 9: Target Confidence Interval Width for Interclass Correlation Coefficients**

Suppose that four radiologists have undergone a new type of training aimed at providing greater consistency of interpretations when assessing MR images of brain. Prior to and after the training, the radiologists will independently assess several sets of images and the proportion of cases each radiologist believes to be abnormal will be recorded. It is expected that the new training will yield an intraclass correlation coefficient (ICC) for interpretation of MR images of approximately 0.8 as compared with an ICC of 0.6 for imaging interpretation prior to the training. What sample size is required to estimate a 90% confidence interval for the ICC of imaging interpretation after training, with a margin of error no wider than 0.05 if the ICC is approximately 0.8?

**Solution:** The Nest function from the ICC package can be used to implement the case for sample size for ICC for particular precision. Note that the best estimate for the anticipated ICC (0.8 in this case) is used in the function call.

```r
library(ICC)
Nest(est.type="hypothetical", ICC=0.8, w=0.05, k=4, alpha = 0.1)
```

Therefore, a sample size of at least 335 images is required to construct a 90% confidence interval no wider than 0.05 for the new ICC value, if it is expected to be close to 0.8.

**Example 10: Target Confidence Interval Width for Proportion—Risk Ratio**

Consider a cohort study in which the investigators aim to determine the relationship of infection in pregnant women by a virus and the development of oligohydramnios during pregnancy as assessed by fetal MRI examinations. Assume that 50% of the study subjects were infected by the virus during pregnancy and 50% were noted. All pregnant women were followed during their pregnancies by fetal MRI examinations performed every 3 months to determine whether they had developed oligohydramnios. If approximately 2.1% of women who had never been infected developed oligohydramnios during their pregnancies, and the relative risk for developing oligohydramnios is estimated to be 11.4, what sample size is required to estimate a 90% confidence interval to a margin of error of no more than 0.025 for the proportion of women who developed oligohydramnios during their pregnancies, after being infected by the virus during pregnancy?

**Solution:** Two proportions, $p_2$, the proportion among the unexposed group, and $p_1$, the proportion in the exposed group, are required to apply a formula.

Given is $p_2 = 0.021$ and $RR = 11.4$ so $p_1 = RRp_2 = 0.23944$, $\alpha = 0.1$ (for a 90% confidence interval) and the width of the interval is supposed to be no wider than 0.05. Conveniently, this becomes very much like a simple confidence interval sample size question.
Sample Size Estimation

Solution: Conveniently, rewriting the code and adding a scaling factor for the balance of the number of controls for every case with the parameter kappa (not to be confused with kappa for reliability) is shown below.

```r
power.roc.test(auc=0.8, sig.level=0.05, power=0.9, kappa=4)
```

## One ROC curve power calculation

```r
ncases = 10.89815
ncontrols = 43.5926
auc = 0.8
sig.level = 0.05
power = 0.9
```

In this case, 11 cases and 44 controls are required. Notice that this is more than the previous total: now a total of 55 images are required but in the previous example only a total of 34 images are required. The possible “advantage” to this is that fewer cases (11 in this case compared to 17 in the last example) are required, which may be appealing if cases are more difficult to be obtained than controls, and may offset the additional controls required. Depending on the rarity of cases, the ratio of cases and controls may be manipulated to obtain a certain power or significance level, when measurements from one group are easier to obtain.

Example 13: Target Power for Diagnostic Test—Area Under the Curve

Consider the previous example (four controls for every case) but now use an AUC value of 0.6, keeping all other parameters the same.

Solution: The solution in this case requires the `power.roc.test` function in the pROC package. This is very similar to the sample size question for kappa and requires the following information.

```r
library(pROC)
power.roc.test(auc=0.6, sig.level=0.05, power=0.9)
```

## One ROC curve power calculation

```r
ncases = 16.6192
ncontrols = 16.6192
auc = 0.6
sig.level = 0.05
power = 0.9
```

Notice that this gives a balanced design: the same number of cases as controls (17 in each case), yielding a total of 34.

Example 12: Target Power for Diagnostic Test—Area Under the Curve

Repeat the previous question imposing a ratio of approximately four controls for every case.
### Example 14: Target Power for Diagnostic Test—Area Under the Curve

Consider further varying the AUC value again from the previous example. If expecting an AUC of 0.99, would you expect the study to require fewer or more cases and controls than the previous example?

**Solution:** In example 13, when AUC = 0.8, more cases and controls were required than if AUC = 0.6. Therefore, if AUC > 0.8, one should expect requiring even fewer cases and controls than the case when AUC = 0.8, and certainly fewer cases than the situation when AUC < 0.6, as this distinction (against the null hypothesis that AUC = 0.5) should be easy to be detected. This is verified with the following code:

```r
data <- data.frame(ncases=1.962138, ncontrols=7.848551, auc=0.99, sig.level=0.05, power=0.9)
```

The power is indeed greater than the required power of 0.9.

### Example 15: Target Power for Reliability

Consider testing the reliability of scoring systems for MRI vs. plain radiographs for evaluating hemophilic joints.11 Considering that the minimum acceptable inter-reader reliability (null hypothesis) is \( \rho_0 = 0.6 \) against the alternative hypothesis that \( \rho_1 = 0.8 \), determine that the minimum number of examinations of knees required to obtain a type 1 error no more than 0.05 and a type 2 error no more than 0.2 using the tables in reference 1.

**Solution:** For four readers, the minimum number of images required is 21.9. This is, of course, rounded up to 22 images. Note that while these tables are thorough, the approximation for sample size from this publication can be used for any combination of null or alternative hypotheses (alternative greater than null) given a maximum type 1 and maximum type 2 error/required power for a given number of raters.

### Example 16: Rule of Thumb Based on the Number of Predictors of an Outcome

One final example somewhat unrelated to the previous concepts, but still in the realm
Sample size for a confidence interval:
\[ n = \left( \frac{2\sigma Z_{a/2}}{w} \right)^2 \]

Sample size for hypothesis test (one-sided alternative)
\[ n = \left( \frac{Z_{a} + Z_{\beta}}{\delta} \right)^2 \sigma^2 \]

Several of the formula are useful for proportions when using \( \sigma^2 = p(1 - p) \)

For confidence intervals, the following factors increase sample size:
- Decreasing the required minimum width of the confidence intervals
- Increasing the confidence level
- Increasing the population variance

For hypothesis tests, the following increase sample size:
- Increasing the required minimum power of the test
- Decreasing the significance level of the test
- Increasing the population variances
- Decreasing the difference between the null and true means

Many variants are possible including
- Using measured variables as a continuous outcome
- Using raw proportions
- Using agreement (kappa), odds ratio, and relative risk for proportions
- Using ICC
- Using AUC

Finally, it was observed that although many questions arise from similar or related formulae, many variants are possible. For this reason, software or consultations with a statistician are recommended to ensure correctness.

### References
2. Hanley JA, Barbara MJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36
16 Meta-analysis
Joseph Beyene, Binod Neupane, and Zelalem F. Negeri

Learning Objectives
- To explain basic statistical and methodological issues related to meta-analysis.
- To explain concepts of effect measures.
- To explain two broad modeling approaches: fixed versus random effects model.
- To discuss heterogeneity and its implications.
- To provide examples of different meta-analytic applications including intervention and diagnostic test accuracy (DTA) reviews.

Introduction
Meta-analysis is a statistical method that combines findings from multiple studies that aim to address the same clinical and scientific question. Meta-analysis methods are used in a wide range of disciplines including medical, social, and educational research. Broadly speaking, meta-analysis may be conducted to achieve two objectives: synthesis and analytic goals. In the context of the synthesis objective, a well-conducted meta-analysis allows for estimation of parameters of interest with better precision and may enable detecting genuine effects with increased power. The analytic goal allows understanding of inconsistencies in effects across studies, exploring potential reasons for the inconsistency, and generating new hypotheses that can be evaluated in a new prospective study.

Meta-analysis has a broad application and combines estimates across studies for different purposes such as evaluation of intervention effect for clinical outcomes or prognosis and assessment of accuracy of a diagnostic test. Both experimental and observational study designs are amenable to meta-analysis.

Experimental studies include the common parallel group randomized controlled trials (RCTs) as well as nonstandard designs such as cluster randomized trials and cross-over trials. Studies assessing the accuracy of a diagnostic test are observational in nature. However, it is possible to have RCTs within a meta-analysis if the purpose of the review is to combine the effects of performing a diagnostic test vs. no such test on some clinical endpoint.

The meta-analysis model choice depends on the nature of the outcome of interest. Meta-analysis can be applied to synthesize effect measures that are calculated from a variety of underlying data structures. Distributional assumptions that reflect the nature of the underlying variables dictate the statistical modeling framework that is used to estimate the relative treatment effects.

Meta-analysis has gained increasing popularity in medical research over the last three decades and radiology is no exception. To show an example of the growth in applications of meta-analytic models in medical imaging, journals falling under the subject category “Radiology, Nuclear Medicine & Medical Imaging” were searched and sorted the journals based on their 5-year impact factor and considered the top seven journals (starting from highest 5-year impact factor: Human Brain Mapping, Neuro Image, Circulation Cardiovascular Imaging, Radiotherapy and Oncology, Radiology, JACC Cardiovascular Imaging, Journal of Nuclear Medicine, Investigative Radiology, Medical Image Analysis, and European Journal of Nuclear Medicine and Molecular Imaging) in our search for articles focusing on meta-analysis.

Fig. 16.1 presents the number of articles published per year until December 31, 2016, in 10 high-impact journals indexed in Ovid Medline in the field of imaging and radiology.
c. Count outcome: effect measure commonly used is rate ratio (RR)

d. Time-to-event outcome: a common effect measure is hazard ratio (HR)

RD, RR, or OR can be used as effect measures when the interest is in evaluating the effect of performing (vs. not performing) a diagnostic test on a binary clinical endpoint. Example 1 illustrates a meta-analysis assessing whether mammogram screening for breast cancer (screening group) compared to no screening (control group) is more effective in preventing breast cancer mortality. Here, the data from the first study in that hypothetical example of meta-analysis are presented in Table 16.1 to illustrate how the estimates of the effect measures can be obtained.

Here, \( n_1 = a + b \) and \( n_2 = c + d \) represent the number of subjects randomized to screening and control groups, respectively, in the study. If \( p_e \) and \( p_c \) are the risks of breast cancer mortality in screening and control groups, respectively, then their estimates are obtained as:

\[
\hat{p}_e = \frac{a}{a + b} = \frac{17}{40} = 0.425 \quad \text{and} \quad \hat{p}_c = \frac{c}{c + d} = \frac{29}{36} = 0.806
\]

(search conducted in May 2017). The figure clearly shows the increasing trend of meta-analysis publications in radiology over time, consistent with what is observed in other medical disciplines.

## Effect Measures

Effect measures represent parameters of interest and vary depending on the type of study as well as the nature of the outcome variable. Here some of the commonly used effect measures in meta-analysis of medical studies are discussed.

### Meta-analysis Evaluating the Effect of a Diagnostic Test on a Clinical Outcome (such as in Intervention Studies)

a. Binary outcome variables: effect measures commonly used include risk difference (RD), risk ratio, also known as relative risk (RR), and odds ratio (OR)

b. Continuous outcome variables: effect measures include mean difference (MD), standardized mean difference (SMD), and ratio-of-means (RoM)

![Fig. 16.1 Number of meta-analyses in radiology published on a yearly basis in the top 7 high-impact factor journals in imaging and radiology during 1991–2016.](image-url)
Meta-analysis

Assessing the Accuracy of a Diagnostic Test

A. Sensitivity (true positive rate) and specificity (false positive rate)
B. Likelihood ratios for: positive test result (LR+) and negative test result (LR−)
C. Diagnostic odds ratio (DOR)

In meta-analysis of diagnostic test accuracy (DTA) studies, the most widely used effect measures are sensitivity and specificity of a test. Sensitivity, also called the true positive rate (TPR), is the proportion of the positive test results among those with a condition versus those without the condition (confirmed often by the existing gold standard or reference test), whereas specificity, also called the true-negative rate (TNR), is the proportion of negative test results among those without the condition. Note that the false-negative rate (FNR) of a test is \(1 - \text{Sensitivity}\), and the false-positive rate (FPR) is \(1 - \text{Specificity}\). Very high sensitivity and specificity (i.e., close to 1) are indicators of an accurate test, whereas sensitivity of 0.50 or lower and specificity of 0.50 or lower indicate that the test is no better than tossing a coin to diagnose a disease or a health condition. Positive likelihood ratio or likelihood ratio for a positive test result (LR+) is the ratio of the proportions of positive test results among those with a condition versus those without the condition. Whereas negative likelihood ratio or likelihood ratio for a negative test result (LR−) is the ratio of the proportions with negative test results among those with

<table>
<thead>
<tr>
<th>Mammogram screening</th>
<th>Breast cancer mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>a = 17</td>
</tr>
<tr>
<td>No</td>
<td>c = 29</td>
</tr>
<tr>
<td>Total</td>
<td>a + c = 46</td>
</tr>
</tbody>
</table>

The risk difference of the screening group and control group is estimated as

\[
\hat{RD} = \hat{p}_e - \hat{p}_c = \frac{a}{a + b} - \frac{c}{c + d} = 0.425 - 0.806 = -0.381
\]

The relative risk in the screening group compared to the control group is

\[
\hat{RR} = \frac{\hat{p}_e}{\hat{p}_c} = \frac{a / (a + b)}{c / (c + d)} = \frac{0.425}{0.806} = 0.527
\]

The odds ratio, which is the ratio of the odds of dying in the screening group to the odds in the control group, is

\[
\hat{OR} = \frac{a/b}{c/d} = \frac{ac}{bd} = \frac{17 \times 7}{29 \times 23} = 0.178
\]

The estimate of \(\hat{RD}\) of 0.381 is interpreted to mean that the breast cancer mortality rate dropped by 38.1% when women were screened for breast cancer. The estimate of \(\hat{RR}\) of 0.527 indicates that the risk of mortality in the screening group is only 52.7% of that of the group of those not screened for breast cancer, a relative risk reduction of 47.3%. The OR is difficult to interpret as it cannot be directly expressed in terms of risk. Here, the odds of mortality are 0.178 in the screened women compared to those who were not screened. However, for a rare outcome, OR approximates RR and may be interpreted in terms of risk.

The estimates of RD, RR, and OR are random quantities and therefore have associated variances, or standard errors (SEs). The point estimates along with the corresponding variances allow proper statistical inference to be made (e.g., 95% confidence intervals for the underlying parameter, \(p\) value for hypothesis tests, etc.). Mathematical details of variances are not provided here.
a condition versus those without the condition. Likelihood ratios have a nice property that allows them to be defined even when a test categorizes a patient’s condition into three or more categories (e.g., no/probable/definite diagnosis for a health condition), whereas sensitivity and specificity can be defined only for binary categories. Diagnostic odds ratio (DOR) is a single measure of overall accuracy of a test. It is the ratio of the odds of positive to negative tests in diseased subjects to the odds of positive to negative tests in nondiseased subjects. Thus, DOR is the ratio of LR+ to LR− when the test result has a binary category. A test with DOR < 1 is worse than randomly calling a case as diseased or nondiseased. In diagnostic accuracy studies, the higher the DOR, the better the test performance is.

Example 2 provides a meta-analysis evaluating the diagnostic performance of ultrasonography (US) and computed tomography (CT) for diagnosing appendicitis in children, partially following the layout of Doria et al.’s meta-analysis. Here, we illustrate how the above test performance measures are calculated using the data from Ang et al. The data is shown as a 2 × 2 contingency table (Table 16.2).

<table>
<thead>
<tr>
<th>Appendicitis diagnosis (by ultrasonography)</th>
<th>Appendicitis (by reference standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>TP = 145</td>
</tr>
<tr>
<td>No</td>
<td>FN = 14</td>
</tr>
<tr>
<td>Total</td>
<td>TP + FN = 159</td>
</tr>
</tbody>
</table>

Abbreviations: TP, true-positives; TN, true-negatives; FP, false-positives; FN, false-negatives; N, number.
Source: Ang et al.

Here, a sensitivity of 0.912 means that the ultrasonography was able to diagnose accurately only 91.2% of patients with appendicitis, while the test failed to identify 8.8% (false-negative rate) of the appendicitis cases. Similarly, a specificity of 0.887 indicates that 88.7% of non-appendicitis cases were accurately identified as not having appendicitis by ultrasonography, while it wrongly diagnosed 11.3% (false-positive rate) of non-appendicitis cases as having appendicitis. The LR+ of 8.07 means that ultrasonography is 8.07 times more likely to identify appendicitis cases among unhealthy subjects than among the healthy ones. Similarly, the LR− of 0.10 indicates that ultrasonography is 0.10 times more likely to identify non-appendicitis cases among diseased subjects than among healthy subjects. The above DOR tells us that the odds of accurate diagnoses (reporting appendicitis among those with appendicitis and failing to report appendicitis
among those without appendicitis) is 81.26 compared to a misdiagnosis.

Note that the above calculated sensitivity and specificity (proportions) and LR+, LR−, and DOR (ratios) are estimates and therefore they are inherently variable. The interested reader can refer to Chapter 3 for additional details on how to calculate variances of these estimates and confidence intervals of those effect measures.

### Fixed and Random Effects Models for Univariate Meta-analysis

Suppose \( k \) studies are included in a meta-analysis. Each study provides a set of aggregate data \((Y_i, \sigma_i^2, i = 1, 2, ..., k)\), where \( Y_i \) is the estimate of the effect measure (e.g., diagnostic log-odds ratio for a diagnostic test) of interest and \( \sigma_i^2 \) is the variance of \( Y_i \). Broadly speaking, there are two models, namely fixed effects (FE) and random effects (RE) models, to represent the extents of variability in these estimates. FE model assumes that all the studies are performed in an identical population and hence are measuring the same “true effect,” \( \mu \), so that any variation in the estimates is only sampling variability (within-study variance). In contrast, the RE model assumes that each study represents a unique population. Therefore, different studies representing different populations have different “true effects,” \( \mu_i \), that are genuinely different beyond sampling variability and are assumed to be distributed around an “average effect,” \( \mu \), with a variance, \( \tau^2 \), called “between-study variance.” Mathematically,

**FE model:** \[ Y_i = \mu + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma_i^2) \]

**RE model:** \[ Y_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma_i^2), \quad \mu_i \sim \text{N}(\mu, \tau^2) \]

Equivalently, the distribution of the estimates \( Y_i \) is \( Y_i \sim \text{N}(\mu, \sigma_i^2) \) under the FE model, and \( Y_i \sim \text{N}(\mu, \sigma_i^2 + \tau^2) \) under the RE model, which dictate the nature of inference about the quantity of interests \( \mu \) and/or \( \tau^2 \). \( \tau^2 \) depends on the types of effect measures used as well as outcome types. Under the inverse variance weighted estimation, the pooled estimate of \( \mu \), interpreted as the “true effect” under the FE model and “average effect” under the RE model, is calculated as

\[ \hat{\mu} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i} \]

where the inverse variance weight for \( i \)th study is: \( w_i = 1/\sigma_i^2 \) for the FE model and \( w_i = 1/(\sigma_i^2 + \tau^2) \) for the RE model. The standard error (SE) of \( \hat{\mu} \), which is the square root of its variance, is obtained as

\[ \text{SE}(\hat{\mu}) = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}} \]

However, we do not know the exact \( \sigma_i^2 \) and \( \tau^2 \) in practice. Therefore, the respective parameters in the weights \( w_i \) are substituted by the corresponding estimates \( s_i^2 \) and \( \hat{\tau}^2 \). The 95% confidence interval (CI) for \( \mu \) is constructed as

95% CI = \( (\hat{\mu} - 1.96 \times \text{SE}(\hat{\mu}), \hat{\mu} + 1.96 \times \text{SE}(\hat{\mu})) \)

The null hypothesis \( H_0: \mu = 0 \) can be assessed through a Z-test as

\[ Z = \frac{\hat{\mu}}{\text{SE}(\hat{\mu})} \sim \text{N}(0, 1) \]

A value of \( Z \leq -1.96 \) or \( Z \geq 1.96 \) indicates statistical significance (i.e., rejects the null hypothesis \( H_0: \mu = 0 \) in favor of the alternative hypothesis \( H_1: \mu \neq 0 \)) at the level of significance \( \alpha = 0.05 \) for a two-tailed test.

In diagnostic test accuracy studies with effect measures such as sensitivity, specificity, or DOR, some transformation of these quantities is required to make their distribution approximately normal, an important assumption made in a meta-analysis as described above, before meta-analyzing them. Suppose that DOR is the effect measure of interest; then log transformed values of the DORs and variances of the log-DOR’s estimates from individual studies are used (logit transformations are often applied if the effect measures are the proportions, e.g., sensitivity and specificity). Let \( Y_i = \log DOR_i \), and \( s_i^2 = \text{variance of log DOR}_i \) in study \( i \). Once the weighted mean \( \hat{\mu} \) of \( Y_i \) is obtained using FE or RE model, the pooled estimate of DOR is obtained by back transformation as, \( \hat{DOR} = \exp(\hat{\mu}) \), and 95% CI of DOR as \( \exp(\hat{\mu} - 1.96 \times \text{SE}(\hat{\mu}), \hat{\mu} + 1.96 \times \text{SE}(\hat{\mu})) \).
DOR has an advantage over using sensitivity and specificity as an effect measure in a meta-analysis in the sense that DOR does not depend on the threshold (as the threshold effect cancels out in each study), whereas sensitivity and specificity do.1 However, DOR also loses the information on the individual sensitivity or specificity, which may sometimes be a more useful measure of test performance.

Note that when there is no heterogeneity (i.e., when $\tau^2 = 0$), the random effects model reduces to a fixed effects model. Also note that the random effect model assigns less weight to larger studies and greater weights to smaller studies than does the fixed effects model; so the combined estimate of effect size from the two models could differ in practice. The choice of fixed or random effects model depends on the presence of different sources of heterogeneity (i.e., clinical and methodological differences across studies) as well as on the extent or degree of heterogeneity estimated or quantified through the data. In the next section, these issues are discussed in more detail.

**Detecting Heterogeneity**

Heterogeneity is an important issue that must be considered carefully in meta-analysis for a clear interpretation of findings. Generalizing a finding to a case in clinical or diagnostic practice would be easier if the results are homogeneous across studies.

The best approach to understanding heterogeneity is by assessing the comparability of studies for important clinical and methodological factors including patient characteristics, types of interventions and comparators, the way outcome variables (endpoints) are defined, time frame, study design, and analysis.

Detecting heterogeneity can also be facilitated using graphical/visual approaches as well as formal statistical tests. For example, visual inspection of the forest plot in Fig. 16.2 might indicate that there is some evidence of heterogeneity as the confidence intervals of the estimates of effects in individual studies are in general not tightly overlapped. Heterogeneity can be formally assessed statistically by testing for it, estimating it, and quantifying it.

### Testing for Heterogeneity

The most widely used statistical test is the Cochran chi-square test, often known as the Q test in the meta-analysis literature. The Q statistic (“total variance” of the estimates of effects in individual studies) is calculated as

$$Q = \sum_{i=1}^{k} w_i (Y_i - \hat{\mu}_{re})^2$$

where $w_i = 1/s_i^2$. When there is no heterogeneity (as assumed under the FE model), $Q\sim X^2_{k-1}$ so
Study | Experimental Events Total | Control Events Total | Risk Ratio | RR 95%-CI W(fixed) W(random)
--- | --- | --- | --- | ---
1 | 17 40 | 29 36 | 0.53 [0.36; 0.78] | 5.7% 10.6%
2 | 70 287 | 104 297 | 0.70 [0.54; 0.90] | 18.9% 13.8%
3 | 20 61 | 23 65 | 0.93 [0.57; 1.51] | 4.1% 8.8%
4 | 6 43 | 12 43 | 0.50 [0.21; 1.21] | 2.2% 4.1%
5 | 6 31 | 5 30 | 1.16 [0.40; 3.40] | 0.9% 3.0%
6 | 0 56 | 5 59 | 0.10 [0.01; 1.69] | 1.0% 0.5%
7 | 8 34 | 14 32 | 0.54 [0.26; 1.11] | 2.7% 5.5%
8 | 10 40 | 5 40 | 2.00 [0.75; 5.33] | 0.9% 3.5%
9 | 208 1204 | 231 1212 | 0.91 [0.77; 1.07] | 42.7% 15.6%
10 | 4 100 | 7 50 | 0.29 [0.09; 0.93] | 1.7% 2.6%
11 | 10 68 | 13 34 | 0.38 [0.19; 0.79] | 3.2% 5.6%
12 | 7 40 | 10 41 | 0.72 [0.30; 1.70] | 1.8% 4.3%
13 | 4 46 | 3 48 | 1.39 [0.33; 5.88] | 0.5% 1.8%
14 | 3 40 | 8 40 | 0.37 [0.11; 1.31] | 1.5% 2.3%
15 | 13 56 | 14 60 | 0.99 [0.51; 1.93] | 2.5% 6.2%
16 | 59 372 | 52 381 | 1.16 [0.82; 1.64] | 9.5% 11.7%

Fixed effect model | 2518 | 2468 | 0.82 [0.74; 0.92] | 100% --
Random effects model | 0.75 [0.61; 0.92] | -- | 100%

**Heterogeneity:** I-squared=49.9%, tau-squared=0.0625, p=0.0121

**Fig. 16.2** Forest plot assessing the effect of mammogram screening on breast cancer mortality. Pooled estimates and study weights under both fixed and random effects models are provided. CI, confidence interval.
that the expected value of $Q$ is $E(Q) = k - 1$ (number of studies minus one). Therefore, an observed $Q$ much greater than $(k - 1)$ indicates the presence of heterogeneity. The $p$ value for the test of heterogeneity using $Q$ test is obtained as: $p = \Pr(X^2_{k-1} \geq Q)$. This test is easy to use in practice but is not a very powerful test, especially when $k$ is small.

### Estimating Heterogeneity

The simplest and most commonly used method of estimating heterogeneity parameter $\tau^2$ is the DerSimonian-Laird method.\(^5\) There are several other methods such as maximum likelihood, restricted maximum likelihood, and Paule-Mandel\(^7\) for estimating $\tau^2$. The DerSimonian-Laird method is computationally simple (does not require an iterative procedure) and has been implemented in popular meta-analytic software/programs such as the Cochrane Review Manager (RevMan), SAS, and some R packages. It performs acceptably when the number of studies is large and $\tau^2$ is small.\(^8,9\) But the method does not perform well compared to other estimators such as maximum likelihood, restricted maximum likelihood, and Paule-Mandel, especially when the number of studies is small, and therefore should not be used. However, even the restricted maximum likelihood approach underestimates $\tau^2$ when the data are sparse,\(^9,11\) so it may not be preferred when the number of false-positives or false-negatives are 0 or very small in most studies. A recent review provides details of different methods of estimating $\tau^2$ and their comparative weaknesses and strengths.\(^12\) The heterogeneity parameter $\tau^2$ depends on the choice of the effect measure as well as on the outcome type and its variability, and therefore, it might not be appropriate to compare the extent of heterogeneity across different outcome types or health problems.

### Quantifying Heterogeneity

In meta-analysis, it is often of interest to quantify the degree of heterogeneity due to genuine differences across studies. A popular index that quantifies heterogeneity above and beyond change is the $I^2$ statistic.\(^13\) This index is based on the $Q$ test and ranges from 0 to 100%.

$$I^2 = \max\left\{0, \frac{Q - (k - 1)}{Q} \times 100\% \right\}$$

where $F$ is set to 0 if $Q < (K - 1)$. Conventionally, an $F$ value of around 25% or less is considered “small,” 25 to 50% is considered “moderate,” 50 to 75% as high, and 75% or more is considered as “substantial” degree of heterogeneity.\(^13\)

### Dealing with Heterogeneity

#### Ignoring Heterogeneity

Sometimes, even when heterogeneity is detected, the investigator may ignore it and employ fixed effect model on practical ground. When the data are sparse (e.g., event rates for a binary outcome are quite rare and/or most trials are small resulting in zero events in one or more intervention arms in several studies or only couples of trials are available for meta-analysis), it may not be appropriate to use the random effects model based on the inverse-variance weighting scheme.\(^14,15\) The fixed effects models with different weighting methods might be more appropriate for such data.

#### Incorporating Heterogeneity

When heterogeneity is detected, the random effects model is most often employed to incorporate the heterogeneity in the estimation of the average effect and/or assessment of whether there exists some non-null average effect. When there are no obvious factors that can be used to explain observed heterogeneity, the random effects assumption allows incorporation of unexplained heterogeneity. As a result, the combined estimate of effect from the random effects model is less precise and, consequently, has lower power to detect non-null average effect compared to the fixed effects model.

#### Understanding/Explaining Heterogeneity

Heterogeneity can be explained by performing meta-analyses on clinically meaningful subgroups or meta-regression by including
When data on potential sources of heterogeneity (e.g., trial characteristics) are available and enough studies are available for meaningful analysis, subgroup meta-analysis assesses whether there exist some effects in different subgroups and whether they are consistent. Meta-regression assesses whether a study level covariate significantly explains the variations in the observed estimates of effects across studies. However, results from such analysis may serve the purpose of hypothesis generation, and should be cautiously interpreted as multiple testing problems are likely in subgroup analysis and ecological fallacy is likely in meta-regression. Appropriate use and interpretation of findings from meta-regression can be found in a recent review.16

## Meta-analysis Examples

In this section the R packages metafor17 (for intervention studies) and mada18 (for diagnostic test accuracy studies) were used to illustrate some of the methodologies described in the previous sections.

### Example 1: Meta-analysis of Intervention Studies (Hypothetical Example)

This example illustrates a meta-analysis of an intervention study with a binary outcome. A researcher wants to perform a meta-analysis with the aim of learning the effectiveness of mammography screening in reducing the mortality from breast cancer for women aged between 39 and 49 years. The researcher has included 16 randomized clinical trials (RCTs) in this meta-analysis. Number of events and total sample size in each group are sufficient to carry out a meta-analysis. Denote the events and the sample size in the first group (screened cases) with “a_i” and “n_{1i},” respectively. Similarly, “c_i” and “n_{2i}” are used to denote the events and the sample size in the second group (control cases). As the outcome is binary and the study design is RCT, one could use relative risk (RR), odds ratio (OR), and risk difference (RD) as effect measures. Here a meta-analysis with RR as an effect measure is illustrated.

### Set working directory and load metafor library

```r
> setwd("~/Documents/Meta-analysis")
```

### Install the metafor package

```r
> install.packages("metafor")
```

### Load the installed metafor package

```r
> library(metafor)
```

### load mammography data

```r
> mammog <- read.csv(file="mammog.csv",header=TRUE)
> attach(mammog)
```

<table>
<thead>
<tr>
<th>Study</th>
<th>ai</th>
<th>n_{1i}</th>
<th>ci</th>
<th>n_{2i}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>40</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>287</td>
<td>104</td>
<td>297</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>61</td>
<td>23</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>43</td>
<td>12</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>31</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>56</td>
<td>5</td>
<td>59</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>34</td>
<td>14</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>40</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>208</td>
<td>1204</td>
<td>231</td>
<td>1212</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>100</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>68</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>40</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>46</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>40</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>56</td>
<td>14</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>59</td>
<td>372</td>
<td>52</td>
<td>381</td>
</tr>
</tbody>
</table>

Calculating the log relative risks and corresponding sampling variances

```r
> effect.size.mammog <- escalc(measure="RR", ai=ai, bi=n_{1i}-ai, ci=ci, di=n_{2i}-ci, data=mammog)
> effect.size.mammog
```
Although the metafor package is used to demonstrate data analysis, the “meta” package is employed to produce the following forest plot (Fig. 16.2). This package can display the weights for the fixed and random effects models in the same forest plot, as shown in Fig. 16.2.

# Install the meta package as follows
>install.packages("meta")

# Then load the installed meta package
>library(meta)

# Fit the model using the meta package
>fit.meta <- metabin(event.e=ai, n.e=n1i, event.c=ci, n.c=n2i, studlab=Study, data=mammog, sm="RR", backtransf=T)
>forest(fit.meta, comb.fixed=T)

# This displays Fig. 16.2

Accordingly, it can be seen that there is some degree of heterogeneity present among the studies as the confidence intervals in the forest plot do not overlap that much.

Results from a fixed effect analysis obtained using the metafor package are shown below.

### Fixed effects model using the metafor package###

>Fixed.effect <- rma(yi, vi, data=effect.size.mammog, method="FE")
>Fixed.effect

Fixed-Effects Model (k = 16)

Test for Heterogeneity:

Q(df = 15) = 29.8873, p-val = 0.0123

Model Results:

<table>
<thead>
<tr>
<th>estimate se</th>
<th>zval</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2116</td>
<td>0.0564</td>
<td>-3.7510</td>
</tr>
<tr>
<td>-0.3222</td>
<td>-0.1010</td>
<td>***</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Interpretation of Results

The results of the fixed effects model are shown above. The important values are shown in bold.

1. It can be seen that as suggested by the forest plot earlier, there is substantial heterogeneity (Q test p value = 0.0123).
2. The average log relative risk is negative in sign, which indicates that the risk of death from breast cancer for women undergoing mammography is lower than the risk for unscreened women. Furthermore, this estimate value is statistically significant (p value = 0.0002).

Random Effects Model

One way to deal with the presence of heterogeneity as in the above data is to use a random effects model. The following R commands allow us to fit a random effects model to the above data.

### Random-effects model using DerSimonian-Liard method ###

```r
>Random.effects <- rma(yi, vi, data=effect.size.mammog, method="DL")
>Random.effects

Random-Effects Model (k = 16; tau^2 estimator: DL)

tau^2 (estimated amount of total heterogeneity): 0.0622

tau (square root of estimated tau^2 value): 0.2494

I^2 (total heterogeneity / total variability): 49.81%

Test for Heterogeneity:
Q(df = 15) = 29.8873, p-val = 0.0123

Model Results:

<table>
<thead>
<tr>
<th>estimate</th>
<th>se</th>
<th>zval</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2847</td>
<td>0.1045</td>
<td>-2.7253</td>
<td>0.0064</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Results

The results for the random effects model are displayed above. The following can be observed from the output.

1. The heterogeneity test statistic (Q) is significant with p value = 0.0123
2. The $I^2$ statistic is about 50%, which indicates a moderate degree of heterogeneity among the true effect measures.
3. The average log relative risk estimate is again negative, telling us that the risk of mortality from breast cancer is lower for the mammography group than for the placebo group. This estimate is again statistically significant (p value = 0.0064).
4. One can also take the exponent of the average log relative risk in order to get the value in the relative risk scale and easily interpret the value. For example $\exp(-0.2847) = 0.75$ indicating that the risk of death from breast cancer for the treated women is on average three-fourths the risk of mortality in the untreated group.

Although the DerSimonian-Laird method is commonly used and currently the only method implemented in RevMan, other methods with better statistical properties are available in the metafor package. An example showing how the heterogeneity parameter is estimated using the well known maximum likelihood method is provided below.

The following R codes generate a well-known maximum likelihood estimate for the heterogeneity parameter.

### Random-effects model using Maximum-likelihood (ML) and metafor package###

```r
>Random.effect.ML <- rma(yi, vi, data=effect.size.mammog, method="ML")
>Random.effects.ML

Random-Effects Model (k = 16; tau^2 estimator: ML)

tau^2 (estimated amount of total heterogeneity): 0.0507

tau (square root of estimated tau^2 value): 0.2252

I^2 (total heterogeneity / total variability): 49.81%

Model Results:

<table>
<thead>
<tr>
<th>estimate</th>
<th>se</th>
<th>zval</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2847</td>
<td>0.1045</td>
<td>-2.7253</td>
<td>0.0064</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

279
### Then read the ‘ChildrenUS’ data from your working directory####

>`ChildrenUS<- read.csv(file="ChildrenUS.csv", header=TRUE)`

>`ChildrenUS`

In this example, there are 23 studies in which appropriate data were available for analysis, as shown below.

#### Example 2: Meta-analysis of Diagnostic Test Accuracy (DTA) Studies

Doria et al.\(^2\) evaluated the diagnostic performance of ultrasonography (US) and computed tomography (CT) for screening appendicitis in children and adults. In this example, in order to demonstrate the methodologies for diagnostic test accuracy studies, the ultrasonography data for screening appendicitis in children, called “Children US,” was used.

For diagnostic test accuracy reviews, the R package mada\(^{19}\) can be used to combine univariate test characteristics (e.g., sensitivity, specificity, diagnostic odds ratio, etc.) or bivariate quantities (jointly modeling sensitivity and specificity). The reader is referred to the chapter on diagnostics for further details and introduction on relevant parameters in the context of diagnostic studies. The following four quantities are sufficient to carry out basic meta-analysis for diagnostic test accuracy studies: true-positives (TP), false-negatives (FN), false-positives (FP), and true-negatives (TN).

```r
>install.packages(mada)
>library(mada)
```

Side-by-side forest plots for sensitivity and specificity based on this dataset are shown in Fig. 16.3. The forest plots of sensitivity and specificity both suggest the absence of heterogeneity among the studies except for the presence of one outlier study for sensitivity.\(^{17}\) However, the assumption of the possible (negative) correlation between sensitivity and specificity enforces the researcher to fit the bivariate random effects model.\(^{20}\)
A bivariate random effects model was fitted to the data to jointly model the pair of sensitivities and specificities across the 23 studies. This model requires estimating a total of five model parameters: two fixed parameters (mean sensitivity and mean specificity) and three random effects (variance of sensitivity, variance of specificity, and covariance between sensitivity and specificity).

```r
# Fit a bivariate random effects model using
# the default restricted maximum likelihood method.
> Random.effects.DTA <- reitsma(ChildrenUS)
> summary(Random.effects.DTA)
```

The results from the bivariate random effects model are shown below.

```
Call: reitsma.default(data = ChildrenUS)

Bivariate diagnostic random-effects meta-analysis

Estimation method: REML

Fixed-effects coefficients

Estimate Std. Error Pr(>|z|) 95%ci.lb 95%ci.ub

tsens.(Intercept) 1.800 0.161 11.187 0.000 1.485 2.115 ***

tfpr.(Intercept) -2.738 0.128 -21.386 0.000 -2.989 -2.487 ***
```

**Fig. 16.3** Forest plots for sensitivity and specificity of ultrasonography in diagnosing appendicitis in children. (Source: Doria et al.)
sensitivity 0.858 -- 0.815 0.892
false pos. rate 0.061 -- 0.048 0.077

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Variance components: between-studies Std. Dev and correlation matrix

 Std. Dev  tsens  tfpr
Tsens  0.604  1.000.
Tfpr  0.425 −0.251  1.000

**Interpretation of Results**

The fixed effects estimates of the two effect measures of interest (sensitivity and specificity) are given both in logit transformed and back-transformed (unit scale) formats.

1. The pooled sensitivity (95% CI) of ultrasound for the diagnosis of appendicitis in children is 85.8% (81.5%, 89.2%). This indicates that the effectiveness of the ultrasound screening to detect appendicitis among children with the condition is on average approximately 86%.

2. On the other hand, the specificity (95% CI) of ultrasound for the examination of appendicitis in children is 93.9% (92.3%, 95.2%). That is, the ability of the US screening to identify children as negative when the child does not have the condition is about 94% on average.

**Remark:** The above procedure could also be repeated for CT to obtain the estimates of sensitivity and specificity for the test. Then, based on the interest of the researcher, it is possible to perform a statistical evaluation of hypotheses to identify the test with better accuracy. However, a different package called DTComPair21 is needed in order to do this and the two tests need to be performed on each subject/trial included in the study (i.e., the test is a paired test). Since the discussion of this topic is beyond the scope of this introductory chapter, the interested reader is advised to refer to the aforementioned package for a thorough discussion.

**Highlights of Key Points**

- In a meta-analysis, measures of effects such as relative risk, odds ratio, or risk difference can be used to express the effect of a diagnostic test on a binary clinical outcome. Measures of accuracy such as sensitivity, specificity, or diagnostic odds ratio can be used to express the accuracy of a diagnostic test.

- Fixed effects and random effects models are two broad assumptions often made in meta-analysis. Fixed effects model assumes that all studies are measuring the same true effect or accuracy, and therefore, there is no important variation in the estimates of effect or accuracy across studies beyond chance. Random effects model assumes that there is a genuine difference or “heterogeneity” in the effects or accuracies across studies that represent different populations.

- The presence of heterogeneity in a meta-analysis impacts the way investigators use the statistical models and interpret the meta-analysis findings. It is often investigated by checking for any differences in the patients’ characteristics and designs/methods across studies, and by visual inspection of the results in graphs. It can be assessed by testing for it, estimating it, and quantifying it using statistical methods.

- When heterogeneity is detected or suspected, random effects model incorporates it in the combined estimate of average effect that would be less precise than when a fixed effect model is employed ignoring such heterogeneity.

- An exploratory analysis in a meta-analysis can be performed through subgroup analysis or meta-regression to determine if certain available factors (e.g., patients’ type or risk of bias in study level) are responsible for heterogeneity, but it often has limited scope.
References

Appendix:  
A Brief Introduction to R

■ Background on R

R is a programming language and software environment for performing statistical data analysis and making graphs. This book uses the R programming language to show how to implement the statistical methods in the chapters on descriptive statistics, diagnostic tests, measurements, statistical inference, regression analysis, sample size, and meta-analysis. All the R code and relevant data in the book are available on the accompanying website, so that readers can run the analyses shown and, with slight modifications, run them on their own data.

■ Why R?

R has all the basic functions needed for any data analysis, but has two key features to recommend it for the clinician doing research in radiology: (1) There is a large collection of add-ons that are available for specialized analyses, such as ROC curve modeling, meta-analysis, and measurement reliability; and (2) R is freely available software and does not require the purchase of an expensive yearly license.

■ How to Get R?

R is available for free download from https://www.r-project.org/. You will be redirected to choose a CRAN (Comprehensive R Archive Network) mirror. These sites all have the same sets of R download files (they mirror the main archive). Pick one closest to home (e.g., http://cran.utstat.utoronto.ca). The next and most important choice is to pick the R version that is compatible with your computer operating system (OSX, Windows, or Linux). From there, the installation follows the standard procedure for that operating system. After you have downloaded and installed R on your system, you will have an icon to click to launch R.

■ RStudio

Although it is possible to work directly in R, we recommend you next download and install another piece of open source software, RStudio (https://www.rstudio.com). This provides a user-friendly interface for R by organizing your code, the output of your analyses, graphs, and help files. Download and install the version for your operating system. When you open RStudio, it automatically opens the version of R on your system.

A more detailed explanation of the installation process for R and RStudio is available in our online supplemental material.

■ Using R

When you first open RStudio, use the menus to create a new file to store your commands (File => New => R Script). This will open a window in the top left of your workspace. You can type in commands here, run them, and save the commands (using the File => Save menus) to run again at a later time. The website for this book contains example scripts for each chapter with all the example code used there.

It is also possible to type in a command in the window at the bottom window (the console), but these commands are not saved. This is useful for running commands that you don’t need to save and rerun later. All of the examples below can be typed directly in to the console, followed by the <RETURN> key. We recommend typing them into the script file you have created and then using the “Run” menu item to run them. This makes it easier to fix the inevitable typographical errors and will give you a record of your work when you

Appendix:
A Brief Introduction to R

■ Background on R

R is a programming language and software environment for performing statistical data analysis and making graphs. This book uses the R programming language to show how to implement the statistical methods in the chapters on descriptive statistics, diagnostic tests, measurements, statistical inference, regression analysis, sample size, and meta-analysis. All the R code and relevant data in the book are available on the accompanying website, so that readers can run the analyses shown and, with slight modifications, run them on their own data.

■ Why R?

R has all the basic functions needed for any data analysis, but has two key features to recommend it for the clinician doing research in radiology: (1) There is a large collection of add-ons that are available for specialized analyses, such as ROC curve modeling, meta-analysis, and measurement reliability; and (2) R is freely available software and does not require the purchase of an expensive yearly license.

■ How to Get R?

R is available for free download from https://www.r-project.org/. You will be redirected to choose a CRAN (Comprehensive R Archive Network) mirror. These sites all have the same sets of R download files (they mirror the main archive). Pick one closest to home (e.g., http://cran.utstat.utoronto.ca). The next and most important choice is to pick the R version that is compatible with your computer operating system (OSX, Windows, or Linux). From there, the installation follows the standard procedure for that operating system. After you have downloaded and installed R on your system, you will have an icon to click to launch R.

■ RStudio

Although it is possible to work directly in R, we recommend you next download and install another piece of open source software, RStudio (https://www.rstudio.com). This provides a user-friendly interface for R by organizing your code, the output of your analyses, graphs, and help files. Download and install the version for your operating system. When you open RStudio, it automatically opens the version of R on your system.

A more detailed explanation of the installation process for R and RStudio is available in our online supplemental material.

■ Using R

When you first open RStudio, use the menus to create a new file to store your commands (File => New => R Script). This will open a window in the top left of your workspace. You can type in commands here, run them, and save the commands (using the File => Save menus) to run again at a later time. The website for this book contains example scripts for each chapter with all the example code used there.

It is also possible to type in a command in the window at the bottom window (the console), but these commands are not saved. This is useful for running commands that you don’t need to save and rerun later. All of the examples below can be typed directly in to the console, followed by the <RETURN> key. We recommend typing them into the script file you have created and then using the “Run” menu item to run them. This makes it easier to fix the inevitable typographical errors and will give you a record of your work when you
save the script. In the sections below, and in the rest of this book, we use > at the beginning of a line to indicate that the rest of the line is what is typed in. For example, the following line means that you type in to the script sqrt(25). Do not type in the “>.” The output is shown directly below input.

```r
> sqrt(25)
[1] 5
```

**An Introduction to R (as a Big Calculator)**

One way to think about R is that it is simply a calculator. Basic mathematical commands are easily accomplished in R as shown below:

```r
> 1+1
[1] 2
> 3*4
[1] 12
> factorial(5)
[1] 120
> 10^2
[1] 100
> log(1)
[1] 0
```

To make a comment in the script to describe what the command does, put the # symbol at the front of the line. R will not run this line, but it will help you understand your thinking when you return to this code later:

```r
# Length of the hypotenuse of a right-angled triangle with the
# short sides having length 6 and 8
> sqrt(6^2 + 8^2)
[1] 10
```

The calculations do not have to include numbers. We can define variables to have certain values, using the assignment operator, “<-” (less than sign, followed by minus sign). We can then make calculations on these variables:

```r
> a <- 3
> b <- 4
> a
[1] 3
> a*b
[1] 12
```

The results of any calculations can be saved into another variable, which we can print out or use later:

```r
> ab <- a*b
> print(ab)
[1] 12
```

There is a shorter way of printing the value: simply type the variable name.

```r
> ab
[1] 12
```

There are two important rules about variable names:

- Variable names must start with a letter and can contain only letters, numbers, and the period (“.”). For example, these are valid names: age, age20, AGE20, Year.2016, Year.when.20.
- R is case sensitive: age20 and AGE20 are different variables.

**Data Structures and Concepts in R**

This book uses several of the variable types available in R: integer, numeric, character, and logical. R decides on the type automatically, based on what is in the variable. The commands below will create a character variable and a numeric variable:

```r
> label <- "Length of lesion"
> Length <- 0.7654
```
Appendix A: A Brief Introduction to R

> SizeData

<table>
<thead>
<tr>
<th>Largest</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.76 A23</td>
</tr>
<tr>
<td>2</td>
<td>0.81 A27</td>
</tr>
<tr>
<td>3</td>
<td>0.65 B03</td>
</tr>
<tr>
<td>4</td>
<td>0.89 B18</td>
</tr>
<tr>
<td>5</td>
<td>0.62 B19</td>
</tr>
</tbody>
</table>

### Importing and Exporting Data

It is rare that a data analyst will type data directly into R as we have done in the examples above. Data are usually collected in database or spreadsheet software, and Chapter 10 has excellent advice on how to manage and collect data so that the resulting files can be easily read into R for analysis. R has many functions to facilitate the data importation process.

Many database and spreadsheet programs allow a database to be exported as a comma-separated variable (CSV) file. This CSV file will have a row for each observation, and columns for each variable, separated by commas. The datasets on our accompanying website all use this format. If we have (a) a file “MRIResults.csv”; (b) in the folder “/users/smith/mriproject”; (c) that has the variable names in the first row (header); (d) with commas (",") separating values within a row; and (d) want to store the data in a data frame called MRI, use the code:

```r
> MRI <- read.table("/users/smith/mriproject/MRIresults.csv", sep="", header=T)
```

Note: Microsoft Windows users, pay special attention to the direction of the (forward) slashes.

Use `help(read.table)` and `help(write.table)` to learn more about reading and writing data with text files.

Often, data have been stored in Microsoft Excel spreadsheets programs (i.e., .xls or .xlsx formats). If there are many worksheets within a single file, it can be tedious to save

---

Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
Doria_Appendix A.indd   287
To create a matrix with a specified number of rows and columns:

```r
> m1 <- matrix(c(5,10,15,20, 35, 30),ncol=3,nrow=2, byrow=T)
> dim(m1)
[1] 2 3
> m1
[,1] [,2] [,3]
[1,] 5 10 15
[2,] 20 25 30
```

To access parts of a matrix or data frame, we can use the row or column number. For example, to access the first row of m1 (and all three columns in row 1), type:

```r
> m1[,1]
[1] 20 35 30
```

To access the second column of SizeData (and all five rows in that column), use the following:

```r
> SizeData[,2]
[1] A23 A27 B03 B18 B19
```

These approaches can be combined, so to access the first and fourth rows of the second column of SizeData, use this command:

```r
> SizeData[c(1, 4), 2]
[1] A23 B18
```

For data frames, there is also the option to use the column name to pick a column. SizeData has columns named “Largest” and “Patient.” We can access the “Patient” column variable using the following:

```r
> SizeData$Patient
[1] A23 A27 B03 B18 B19
```

Each chapter of statistical methods uses and explains functions specific to the chapter, but there are some basics that reappear throughout the book that we demonstrate below. To assign the sequence 1, 2, 3, . . . , 7 to b, use the “:” operator; then to find out the length of the vector b, use the length() function.

```r
> b <- 1:7
> b
[1] 1 2 3 4 5 6 7
> length(b)
[1] 7
```

To make logical comparisons, use the >, >, ==, >5, <=, and ! (not equal to) operators. These create a special type called a logical variable.

```r
> b == 2
FALSE TRUE FALSE FALSE FALSE FALSE FALSE
> above5 <- b > 5
> class(above5)
[1] “logical”
```

A logical value acts like a 0 for FALSE and 1 for TRUE in arithmetic, so to count how many values are larger than 5, we can use:

```r
> howmany <- sum(above5)
> howmany
[1] 2
```
To Learn More about R

After a few hours of trying out the kinds of commands in this chapter and examining the results, you should have a grasp of many of the basics in R. The statistical chapters of this book aim to provide a good foundation in many of the kinds of analyses useful to a researcher in radiology. However, it can take many years to learn how to take advantage of the richness and power of R. To learn from some of the experts, we refer the interested reader to the following textbooks:

Statistical Data Analysis in R
- Dalgaard P. Introductory Statistics with R. New York: Springer; 2008

Advanced Programming in R
- Spector P. Data Manipulation with R. New York: Springer; 2008
- Wickham H. Advanced R. Boca Raton, FL: CRC Press; 2015

Packages in R

The standard R installation includes a core set of basic functions for data analysis. However, as of the writing of this book (March 2016), R’s functionality can be extended by downloading and installing more than 8000 user-contributed packages from CRAN. One driving force behind R’s current and growing popularity is the availability of R packages for virtually any statistical task.

As an example, suppose you are interested in ROC analysis in R and a Google search identifies pROC as the package that you need. With an active Internet connection, you download and install the package as follows:

```
> install.packages("pROC")
```

This will give you feedback on the location from which the package is being downloaded, success of the process, and a number of other details we do no need to be concerned with here. Once the package is downloaded, it is on your computer permanently, but needs to be called within the R session before you can use it, using the library command. The help command will open up the help file for the library, information that is also available through the “packages” tab on the lower right window in RStudio.

```
> library(pROC)
> help("pROC")
```
Index

Note: Italic b, f, and t indicate box, figure, and table, respectively.

A
Abbreviation use, 182–183
Abstracts
presentation in scientific meetings, 192
section of manuscript, 185, 186f
Academia (academic radiology)
career in, reasons for, 210–211
clinician-scientists in, 219
components/mission in, 211
first five years in, 212–219
grants and grant writing in, 210–215, 214t–217t
ingredients for success in, 223
leadership in, 221–223, 222t
mentor vs. supervisor in, 219–220
mid-cycle review in, 218
private practice vs., 210
promotion requirements in, 215–218
publishing and, 174, 212–213
research as success tool in, 210–224
time management in, 220–221
Academic detailing, 131
Academic medical centers (AMCs), 211. See also Academia
Academy of Radiology Leadership and Management, 222t
Accuracy, 3–4, 4f, 11t, 30–32. See also Validity
meta-analysis of, 269, 271–273, 280–282, 281f
sample size estimation and, 257
Acknowledgements, 190
Adjustment, 95
multiple, 96t
regression, 97–98
simple, 96t
Adverse events
in economic evaluation, 155, 161, 165–166, 167
in results section, 183
risk estimates of, 140
time horizon for, 167
Allocation bias, 107, 108
Allocation concealment, 108–109
Alternative hypothesis, 227–229, 229t
AMED database, 119
American College of Radiology (ACR) Leadership Institute, 222t
American Roentgen Ray Society (ARRS) Clinician Educator Development Program, 214t
Scholar Award, 216t
Analysis of variance (ANOVA), 63, 230t, 239–241
Analysis (discussion) section, of manuscript, 179, 181t, 183–184, 184t
Analytic goal, 117
Analytic research, 2–3, 3f
Angiography, CT vs. conventional, 149t, 147–148, 147t, 148t
ANOVA. See Analysis of variance
Appendicitis
CT for bias in, 51
economic evaluation of, 168–169
low-dose vs. standard, 39–47, 39t, 41f, 42f
meta-analysis of, 272–273, 280–282
ROC curve for, 40–43, 42f
systematic review on, 119, 120t–121t
ultrasound for bias in, 51
economic evaluation of, 168–169
meta-analysis of, 272–273, 272t, 280–282, 281f
predictive criterion validity of, 74, 74f
sensitivity and specificity of, 32–35, 32t, 34t, 272–273, 272t, 280–282, 281f
systematic review on, 119, 120t–121t
Appendix, in manuscript, 182t
Area under the curve (AUC), 43–44, 44f
responsiveness in, 83
sample size estimation for, 259–260, 265–266
Arithmetic mean, 17–19, 20f
Arnold–Hilgartner radiographic scale, 56, 56f
Ascertainment bias, 51, 105
Assessing the Methodological Quality of Systematic Reviews (AMSTAR), 127
Association between variables
decision tree on, 251f
measures of, 26–28
Association of University Radiologists (AUR), 214f
Asthma Quality of Life Questionnaire, 166
Attrition bias, 122–123
Audit, 131
Automated updates (auto alerts), 177–178, 183
Availability heuristic, 141
Average (mean), 17–19, 20f

B
Bar graphs, 13–15, 14f–15f
Belmont Report, 196
Best case/worst case, 96f
Bias. See also Confounding allocation, 107, 108
ascertainment, 51, 105
attrition, 122–123
case-control studies and, 10
cohort studies and, 7
decision making and, 141–142
definition of, 7
detection, 122–123
in diagnostic tests, 49–51
in economic evaluation, 162
instrument, 61
lead time, 50, 50f
length time, 50, 51f
measurement, 58–62
observational research and, 92, 93t–94t
observer, 61
performance, 93t, 122–123
publication, 94f, 105, 127, 128, 191
randomized controlled trials and, 9, 10,
104–106, 108
research design and, 10
sampling, 51, 93t, 104–105
selection, 51, 93t, 104–105, 122–123, 127
strategies to reduce, 61–62
subject, 61
systematic reviews and, 119–124, 122f,
127–128
verification, 50–51, 93t, 105
workup, 51, 93t, 105
Bias modeling, 127–128
Bimodal histogram, 15–17, 16f
Binary data, 229–236
multiple groups with, 230t, 234–235
one group with, 230–232, 230t
paired observations with, 230t, 235–236,
236t
two groups with, 230t, 232–234, 232t
Bland–Altman plot, 64f
Blinding, 61–62, 71, 200
heterogeneity from lack of, 274
in randomized controlled trials, 61–62,
105, 109, 114
in systematic reviews, 119–122
Block randomization, 108
Bone metastases, scan specificity for
infections with, 48–49, 48f
Boolean operators, 177
Bowel injury, CT in, 45–47, 46f
Boxplot, 22–24, 23f–24f
Breast cancer, mammogram for
digital vs. film, 43–44, 44f
meta-analysis of, 270–271, 271t,
277–280
British Medical Journal checklist, 125
Budget, 2, 204–207
common omissions, 205
definition of, 204
expense categories and eligibility, 205
in-kind contributions in, 205–207
justification for, 207
reporting milestones and deliverables, 207
template, 205, 205f–206f

C
Calibration, 59, 61, 71
Campbell & Cochrane Economics Methods Group, 125
Canadian Institutes of Health Research (CIHR), 211, 213–214, 214t
Canadian Task Force on Preventive Health Care, 127
CANCERLIT database, 119
Cardiac tests, noninvasive, retrospective cohort study of, 99–102
Case-control design, 1, 4, 5, 8–9, 8f, 10, 11t, 88, 91–92
confounding in, 91, 94–95
matching in, 91, 94, 95–97
odds ratio in, 98
time points in, 91, 91f
Case report forms (CRFs), 199–200
Case reports, 88
Case series, 88
Categorical variables, 250–252, 251f
CBA. See Cost–benefit analysis
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA. See Cost-effectiveness analysis</td>
<td></td>
</tr>
<tr>
<td>Centrality, measures of, 17–21</td>
<td></td>
</tr>
<tr>
<td>Cervical cancer, decision analysis in, 149–150</td>
<td></td>
</tr>
<tr>
<td>Chance node, 143–144</td>
<td></td>
</tr>
<tr>
<td>Channeling bias, 93t</td>
<td></td>
</tr>
<tr>
<td>Chi-squared test</td>
<td></td>
</tr>
<tr>
<td>Cochran (Q test), 274–276, 278, 280</td>
<td></td>
</tr>
<tr>
<td>Kruskal–Wallis, 230t, 239–241</td>
<td></td>
</tr>
<tr>
<td>McNemar’s, 46, 230t, 236</td>
<td></td>
</tr>
<tr>
<td>multiple regression analysis, 254</td>
<td></td>
</tr>
<tr>
<td>Pearson’s, 230t, 235</td>
<td></td>
</tr>
<tr>
<td>Chronology bias, 93t</td>
<td></td>
</tr>
<tr>
<td>CINAHL (nursing database), 119</td>
<td></td>
</tr>
<tr>
<td>Citation bias, 94t</td>
<td></td>
</tr>
<tr>
<td>Clinical decision aids, 150</td>
<td></td>
</tr>
<tr>
<td>Clinical decisions, 138–142. See also Decision analysis</td>
<td></td>
</tr>
<tr>
<td>bias and, 141–142</td>
<td></td>
</tr>
<tr>
<td>heuristics and, 141</td>
<td></td>
</tr>
<tr>
<td>improving, 142</td>
<td></td>
</tr>
<tr>
<td>lack of knowledge for, 139</td>
<td></td>
</tr>
<tr>
<td>risk estimation difficulties in, 139–140</td>
<td></td>
</tr>
<tr>
<td>uncertainty in, 140–141</td>
<td></td>
</tr>
<tr>
<td>values and, 142</td>
<td></td>
</tr>
<tr>
<td>Clinical outcome, meta-analysis of, 270–271, 271t</td>
<td></td>
</tr>
<tr>
<td>Clinical practice guidelines, development of, 126–127</td>
<td></td>
</tr>
<tr>
<td>Clinical prediction rules, 150</td>
<td></td>
</tr>
<tr>
<td>Clinician-scientists, 219</td>
<td></td>
</tr>
<tr>
<td>Cluster bias, 94t</td>
<td></td>
</tr>
<tr>
<td>Clustered data, 202</td>
<td></td>
</tr>
<tr>
<td>Cluster randomized trials, 269</td>
<td></td>
</tr>
<tr>
<td>Cluster sampling design, 236</td>
<td></td>
</tr>
<tr>
<td>CMA. See Cost-minimization analysis</td>
<td></td>
</tr>
<tr>
<td>Cochran chi-square test (Q test), 274–276, 278, 280</td>
<td></td>
</tr>
<tr>
<td>Cochrane Collaboration, 125–126, 133</td>
<td></td>
</tr>
<tr>
<td>Cochrane Controlled Clinical Trials Register (CCTR), 119</td>
<td></td>
</tr>
<tr>
<td>Cochrane Database of Systematic Reviews (CDSR), 126</td>
<td></td>
</tr>
<tr>
<td>Cochrane Review Manager (RevMan), 276, 279</td>
<td></td>
</tr>
<tr>
<td>Code of Federal Regulation (CFR), 196</td>
<td></td>
</tr>
<tr>
<td>Cohen’s kappa coefficient, 68–71, 260, 263–264</td>
<td></td>
</tr>
<tr>
<td>Cohort design, 1, 4, 5–7, 6f, 10, 11t, 88, 89–91, 90f</td>
<td></td>
</tr>
<tr>
<td>economic evaluation in, 161 example of, 99–102</td>
<td></td>
</tr>
<tr>
<td>prospective, 6–7, 6f, 11t, 89–91, 90f relative risk in, 98</td>
<td></td>
</tr>
<tr>
<td>retrospective, 6, 6f, 7, 11t, 89–91, 90f, 99–102</td>
<td></td>
</tr>
<tr>
<td>Co-investigators, 175</td>
<td></td>
</tr>
<tr>
<td>Commentary, 117</td>
<td></td>
</tr>
<tr>
<td>Computed tomography (CT)</td>
<td></td>
</tr>
<tr>
<td>for appendicitis</td>
<td></td>
</tr>
<tr>
<td>bias in, 51</td>
<td></td>
</tr>
<tr>
<td>economic evaluation of, 168–169</td>
<td></td>
</tr>
<tr>
<td>low-dose vs. standard, 39–47, 39t, 41f, 42f</td>
<td></td>
</tr>
<tr>
<td>meta-analysis of, 272–273, 280–282</td>
<td></td>
</tr>
<tr>
<td>ROC curve for, 40–43, 42f</td>
<td></td>
</tr>
<tr>
<td>systematic review on, 119, 120t–121t</td>
<td></td>
</tr>
<tr>
<td>comparison of findings from two readers, 45–47, 46t</td>
<td></td>
</tr>
<tr>
<td>MRI vs. economic evaluation of, 153–154</td>
<td></td>
</tr>
<tr>
<td>Computed tomography angiography, vs.</td>
<td></td>
</tr>
<tr>
<td>conventional, 149t, 147–148, 147f, 148t</td>
<td></td>
</tr>
<tr>
<td>Concept map, 176</td>
<td></td>
</tr>
<tr>
<td>Concurrent validity, 73, 75–76, 75f, 76f</td>
<td></td>
</tr>
<tr>
<td>Conditioning, 31, 95</td>
<td></td>
</tr>
<tr>
<td>Confidence interval, 32–35, 226–227, 227f, 248</td>
<td></td>
</tr>
<tr>
<td>sample size estimation based on, 257–259, 258f</td>
<td></td>
</tr>
<tr>
<td>target CI width for count outcome, 261</td>
<td></td>
</tr>
<tr>
<td>target CI width for difference in proportion, 262</td>
<td></td>
</tr>
<tr>
<td>target CI width for ICC, 264</td>
<td></td>
</tr>
<tr>
<td>target CI width for kappa, 263–264</td>
<td></td>
</tr>
<tr>
<td>target CI width for proportion outcome, 262, 264–265</td>
<td></td>
</tr>
<tr>
<td>Confidentiality, data, 198–199, 201–202</td>
<td></td>
</tr>
<tr>
<td>Confirmatory bias, 141–142</td>
<td></td>
</tr>
<tr>
<td>Conflicts of interest statement, 182t</td>
<td></td>
</tr>
<tr>
<td>Confounding</td>
<td></td>
</tr>
<tr>
<td>controlling, 95–98, 96t</td>
<td></td>
</tr>
<tr>
<td>definition of, 92, 105</td>
<td></td>
</tr>
<tr>
<td>by indication, 9, 93–95, 93t, 105–106, 106f in measurement error, 59</td>
<td></td>
</tr>
<tr>
<td>in observational design, 91, 92–97, 95f in randomized controlled trials, 105–106</td>
<td></td>
</tr>
<tr>
<td>Consent, informed, 181, 196–198</td>
<td></td>
</tr>
<tr>
<td>Consequences, in economic evaluation, 154</td>
<td></td>
</tr>
<tr>
<td>Conservative estimate, 262</td>
<td></td>
</tr>
<tr>
<td>Consistency (internal reliability), 62–63</td>
<td></td>
</tr>
<tr>
<td>CONSORT statement, 112–114, 113f, 182</td>
<td></td>
</tr>
<tr>
<td>Construct validity, 76–77, 77f</td>
<td></td>
</tr>
<tr>
<td>Consumption, in economic evaluation, 154</td>
<td></td>
</tr>
<tr>
<td>Contamination bias, 94t</td>
<td></td>
</tr>
<tr>
<td>Content validity, 72–73, 74f</td>
<td></td>
</tr>
<tr>
<td>Context bias, 94t</td>
<td></td>
</tr>
<tr>
<td>Continuous data, 57, 58f, 229</td>
<td></td>
</tr>
<tr>
<td>multiple groups with, 230t, 239–241, 239f one group with, 230t, 236–237</td>
<td></td>
</tr>
<tr>
<td>paired observations with, 230t, 241–243</td>
<td></td>
</tr>
<tr>
<td>two groups with, 230t, 238–239</td>
<td></td>
</tr>
</tbody>
</table>
Continuous outcomes
in meta-analysis, 270–271
sample size estimation for, 261–262
Contributorship, 178–179, 190
Controls, in case-control studies, 8–9, 8f
Convergent construct validity, 76–77, 77f
Copayments, 158, 159t
Copyright, transfer of, 182t
Coronary artery, Doppler ultrasound of, 82–83
Correlated data, 202
Correlational analysis, 83, 260
Correlation coefficients, 68, 248–249. See also specific measures
Cost–benefit analysis (CBA), 125, 155, 155t, 156
Cost-consequence table, 167, 168t
Cost-effectiveness analysis (CEA), 125, 150–151, 155–157, 155t, 167–168
Cost-effectiveness plane, 156–157, 157f
Costing, 158–163
Costing table, 162, 163t
Cost item identification, 158–160, 159t, 160t
Cost item measurement, 158, 160–162
Cost-minimization analysis (CMA), 155, 155t, 156
Cost–utility analysis (CUA), 125, 155, 155t, 156
Cost valuation, 158, 162–163, 163t, 164t
Count outcome
in meta-analysis, 270–271
sample size estimation for, 261
Covariate, in linear regression, 245
Cover letter to editor, 182t, 188–190, 189f
Covey’s seven habits, 221
Creating and Optimizing the Research Enterprise Workshop (CORE), 222
Criterion validity, 73–76, 74f–76f
concurrent, 73, 75–76, 75f, 76f
predictive, 73–74, 74f
Cronbach, Lee, 62
Cronbach’s α (alpha), 62–63
Cross-over, in clinical trials, 110, 269
Cross-sectional design, 1, 4–5, 5f, 10, 11t, 88–89, 89f
CUA. See Cost–utility analysis

D
Data
analysis and interpretation of, 229–243, 230t
binary, 229–236
multiple groups with, 230t, 234–235
one group with, 230–232, 230t
paired observations with, 230t, 235–236, 236t
two groups with, 230t, 232–234, 232t
clustered, 202
continuous, 57, 58f, 229
multiple groups with, 230t, 239–241, 239f
one group with, 230t, 236–237
paired observations with, 230t, 241–243
two groups with, 230t, 238–239
correlated, 202
descriptive statistics for, 13–29
in economic evaluation, synthesis of, 170–173
electronic, 200
interval (discrete), 57, 57f
multiple observations per participant, 202, 203f
nominal, 56, 229
ordinal, 1, 2f
qualitative, 13
quality control, 204
quantitative, 13
secondary, 1, 2f
in systematic reviews
extraction of, 123
synthesis of, 123–126
types of, 56–58, 59f
Databases, 119, 176–178, 202–204
Data coding, 201
Data collection, 199–204
Data collection forms (DCF), 199–200
Data confidentiality, 198–199, 201–202
Data dictionary, 199
Data elements
defining, 199
standardizing, 200–201, 201f
Data flow, 204
Data formats, 200–201
Data management, 199–204
Data security, 202, 204
Data storage, 202
Data transfer, 205
Decision aids, 150
Decision analysis, 1, 138–152
decision features in, 142
definition of, 138
economic evaluation in, 125, 168
example from published literature, 149–150
improving clinical decision making in, 142
key aspects of, 146–148
options in, 146, 149
outcomes in, 147, 150
parameters in, 146–147, 149
patient characteristics in, 138–139, 142
perspective in, 144, 146, 149
probability in, 138–142, 146–147, 149
sensitivity analysis in, 147
time frame for, 147, 150
values in, 142, 147, 150
Decision node, 143–144
Decision tree, 142–146, 143f
analysis of, 144–146
in assessing relationship between variables, 250f
collection of, 143–144
economic evaluation in, 168
expected value in, 144–145, 149t
explicit components of, 142
implementation of, 147, 150
implicit components of, 142
limitations of, 146
nongraphical aspects of, 144
rationality in, 146
“rolling back,” 145, 149t
simple, using diagnostic testing strategy, 147–148, 147f, 148t
Dependent variable, 245
Depression Scale, 75
DerSimonian–Laird method, for estimating heterogeneity, 276, 279
Description section, of manuscript, 179, 180t, 181–183
Descriptive research, 2, 3f
Descriptive statistics, 13–29
detection bias, 122–123
Deviation, 25–26, 25t
Diagnostic odds ratio (DOR), 271–274
Diagnostic-review bias, 94t
Diagnostic tests, 30–53
accuracy of, 3–4, 4f, 11t, 30–32
meta-analysis of, 269, 271–273, 280–282, 281f
sample size estimation and, 257
bias in, 49–51
comparing results from two, 44–47
data synthesis of studies of, 123–124, 124f
decision tree for, 147–148, 147f, 148t
design of, 3–4
evaluation of, 3–4, 4f
heterogeneity in studies of, 274
likelihood ratios in, 35–39, 37f
reference standard for, 3, 5, 30–31
reliability of, 3–4, 4f, 10
research designs for, 10–12
ROC curve for, 40–43, 42f, 44f
sample size calculations and, 257, 259–260, 265–266
validity of, 3–4, 10, 54–56, 71–77
Diffusion of knowledge, 129
Digital Imaging and Communications in Medicine (DICOM), 198–199
Digital mammography, vs. film, 43–44, 44f
Direct health care costs, 158, 159t
Direct patient costs, 158–159, 159t
Discrete (interval) data, 57, 57f
Discriminant validity, 76, 77, 78f
Discussion section, of manuscript, 179, 181t, 183–184, 184t
Disease spectrum bias, 93t
Dissemination of knowledge, 129
Divergent validity, 76, 77, 78f
Documentation ethics, 197
research (manuscript), 179–190, 180t–181t
Doppler ultrasound, of coronary artery, 82–83
Dropouts, in clinical trials, 110
DTComPair, 282
Dual energy X-ray absorptiometry (DEXA), measurement error in, 58–61
Dummy variable coding, 250–252
E
Economic evaluation, 1, 153–173
analytic approaches to, 155–157, 155t
bias in, 162
concepts in, 154–172
cost-effectiveness acceptability curve in, 171, 171f
cost-effectiveness analysis in, 125, 150–151, 155t, 159t, 167–168
cost-effectiveness plane in, 156–157, 157f
costing in, 158–163
costing table in, 162, 163t
cost item identification in, 158–160, 159t, 160t
cost item measurement in, 158, 160–162
cost valuation in, 158, 162–163, 163t, 164t
data synthesis of, 125, 170–173
decision analysis in, 125, 168
defining, 154–157, 155f
goals of, 154
in health care, reasons for, 153
knowledge translation of, 170–173
mean costs per patient in, 162, 164t
outcome measurement in, 165–166, 165t
quality of life in, 166
in radiology, reasons for, 153–154
reporting, 170–173
research design for, 160–161
research question in, 157, 158b
scatterplot in, 169, 170f
sensitivity analysis in, 168–170, 170f
study perspective in, 163–165, 164f
time horizon in, 167
tornado diagram in, 169, 169f
uncertainty in, 168–170
Educational materials, 131
Educational meetings, 131
Educational outreach, 131
Effect measures, in meta-analysis, 270–273
Effect modifier, 98, 99f
Effect size statistics, 78–80
Electronic data, 200
Electronic data entry forms, 202–204
Electronic submission, of manuscript, 191
Eligibility, for controlled trials, 9, 107
EMBASE, 119, 177
Empirical method, 44
End-of-grant knowledge translation, 129–130
Envelopes, for random assignment, 109
EQ-5D, 166
Equipoise, 87, 110
Error, 58–62
linear regression, 247
margin of, 258
random, 59, 60f
sources of, 58–61, 60f
standard, 79, 227
of effect measures, 271
in meta-analysis, 271, 273
in one group with binary data, 230–231
in one group with continuous data, 237
in regression analysis, 247, 248, 252
of risk difference, 232
in sample size estimation, 258, 259, 262, 264
in two groups with continuous data, 238
systematic, 59–61, 60f
type 1, 229, 229t
type 2, 229, 229t
Ethics, 196–199
data confidentiality, 198–199
documentation, 197
guidelines, 196–197
informed consent, 181, 196–198
Institutional Review Board, 197
retrospective vs. prospective studies, 198
European Organization for Research and Treatment of Cancer Quality of Life (EORTC-QoL), 55, 166
Evidence-based imaging (EBI), 116–117.
See also Systematic reviews
interpretation of results, 126–127
Users' Guide to Medical Literature, 126
Evidence-based medicine (EBM), 116
Exact binomial test, 231
Executive coach, 222
Expected value, 19, 144–145, 149t
Experimental research, 1, 2–3, 2f, 3f, 9–10, 104–115. See also Randomized controlled trials
External reliability, 63
External responsiveness, 81–83, 81f
Extrapolation of model, 248
F
Face validity, 72, 72f, 73f
False negative, 31, 148, 271
False positive, 31, 148, 271
Fatty tumors, MRI specificity for, 47–48, 47f
Feedback, 131
Figures, in manuscript, 182t, 187–188, 187f
Film mammography, digital vs., 43–44, 44f
Financial disclosure, 182t
FINER characteristics, of research question, 2
Fisher, Sir Ronald, 178
Fisher's exact test, 230t, 233–234
Five-number summary, 24
Five Ps, for academic promotion, 215–216
Fixed effects (FE) model, 273–274, 278–279, 281–282
Follow-up bias, 93t
Food and Drug Administration (FDA), 196
Forest plots, 133, 274, 275f, 278–279, 280, 281f
Franklin, Rosalind, 179
Frequency matching, 96
F-test, 249–252
Funding. See Budget
G
Galley proofs, review of, 192
Generalized linear models (GLMs), 245
GERRAF Award, 216t
Ghost authorship, 179
Global F-test, 249–252
Glossary (data dictionary), 199
Gold standard, 3. See also Reference standard
Good Clinical Practice (GCP), 196–197
Goodness of fit, 83, 252, 253f
Grades of Recommendation, Assessment, Development, and Evaluation (GRADE), 125, 127
Grants and grant writing, 210–215, 214t–217t
Green Depression Inventory, 75
Guideline for Good Clinical Practice (GCP), 196–197
Index

Harm, research design to study, 10
Hazard ratio (HR), 257f, 270–271
Health Assessment Questionnaire, 166
Health care resources, in economic evaluation, 154
Health-related quality of life (HRQoL), 54, 55, 71, 166
Health Utilities Index, 166
"Healthy volunteer bias," 93t
Helsinki Declaration, 196
Hemophilic arthropathy, MRI of, 54–55, 76, 76f–78f, 77
Heterogeneity, 274–277
causes of, 274
dealing with, 276–277
definition of, 274
detection of, 274–276, 275f
estimating, 276
examples of, 278, 280
ignoring, 276
incorporating, 276
quantifying, 276
testing for, 274–276, 278, 280
understanding/explaining, 276–277
Heuristics, 141
Hippocampus, reliability of MRI, 63–68, 63f–65f
Histograms, 13–17, 15f–17f
boxplot vs., 22, 23f–24f
extreme values, mean, and median on, 20, 20f
skewed, 17, 17f, 22, 23f
Hodgkin lymphoma, PET/CT in, 161
Holm method, 240–241
Honorary authorship, 179
"How I Do It" series, 117
"Hypothesis-generating studies," 88
Hypothesis testing, 227–229
examples
in CT by two readers, 45–47
in CT for appendicitis, 45
in ultrasound for appendicitis, 32–35
in ultrasound for liver malignancies, 227–229, 229f
for multiple groups with binary data, 234–235
for multiple groups with continuous data, 239–241
for one group with binary data, 230–232
for one group with continuous data, 236–237
for paired observations with binary data, 235–236
for paired observations with continuous data, 241–243
sample size estimation based on, 257, 259, 260f
for two groups with binary data, 232–234
for two groups with continuous data, 238–239
ICER. See Incremental cost per unit of effect
Illustrations, in manuscript, 182t, 187–188, 187f
Imperfect reference standard, 51
Imperfect standard bias, 94t
Implementation research, 129
Incorporation bias, 94t
Incremental cost per unit of effect (ICER), 167–170, 168f
Independent Ethics Committee (IEC), 197
Independent variable, 245, 246f
Index test
in case-control design, 8, 10
in cohort design, 6–7, 6f, 10
in cross-section design, 5, 5f, 10
Indirect costs, 158, 159–160, 160f
111Indium WBC scan, specificity of, 48–49, 48f
Inference, 225–244, 248
Information bias, 93
Informed consent, 181, 196–198
Informed Consent Form (ICF), 197–198
In-kind contributions, 205–207
Institutional Review Board (IRB), 197–198
Instrument bias, 61
Instrument variability, 59, 60f
Intangible costs, 160
Integrated knowledge translation, 129
Intention-to-treat analysis, 110–112, 111f
Internal reliability, 62–63
Internal responsiveness, 78–80, 79f
International Committee of Medical Journal Editors (ICJME), 178–179
International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), 196
Interquartile range, 26
Inter-rater reliability, 3, 63–68
Interval data, 57, 57f
Interval estimation, 225–227
Intervention
meta-analysis of studies, 270–271, 271t, 277–280
in randomized controlled trials, 109–110, 114
Interviewer bias, 93t
Intraclass correlation coefficient, 63, 64–68, 260, 264
297
Liver nodules, kappa coefficient and findings of, 69 ultrasound for malignancies, statistical model of, 225–229, 227f, 229t
Local opinion leaders, 131 Logistic regression, 252–255 Log of the odds, 254 Longitudinal design, 236 Loss to follow-up bias, 93t Lung tumors, MRI responsiveness for, 77–78 Magnetic resonance imaging (MRI) of cervical cancer, decision analysis of, 149–150 economic evaluation of, 153–154 of fatty tumors, sensitivity of, 47–48, 47f of hemophilic arthropathy, 54–55, 76, 76f–78f, 77 of hippocampus, reliability of, 63–64, 63f, 64f, 65f of lung tumor volume, responsiveness of, 77–78 of rabbit knees, responsiveness of, 78, 79f, 80, 82, 82f Mammogram film vs. digital, 43–44, 44f meta-analysis of, 270–271, 271t, 277–280 Mann–Whitney test, 230t, 238–239 MANTIS database, 119 Manuscript, 179–190, 180t–181t abbreviation use in, 182–183 abstract section of, 185, 186t acknowledgement in, 190 body of, 182t cover letter for, 182t, 188–190, 189f description section of, 179, 180t, 181–183 discussion section of, 179, 181t, 183–184, 184t electronic submission of, 191 figures or illustrations in, 182t, 187–188, 187f introduction section of, 184, 185t length of sections, 183, 184 likelihood of publishing, tips to improve, 193–194 organization of, 182t peer review of, 192 references in, 185 review by authors, 191 review of galley proofs, 192 submission for publication, 191–192 submission requirements for, 190 supplemental material in, 182t tables in, 187 testing (results) section of, 179, 180t, 183 Linear mixed effect models, 63 Linear regression diagnostics for, 252, 253f inference in, 248 multiple, 249–252, 251f simple, 245–249, 246f Literature search, 1, 119, 121f, 176–178, 183

K
Kahneman, Daniel, 141 Kappa coefficient, 68–71, 260, 263–264 K awards (NIH), 214, 216t Knee degeneration
- analysis and interpretation of data on, 229–243, 230t multiple linear regression model of, 250–251, 251f multiple logistic regression model of, 254–255 simple linear regression model of, 245, 246f Knowledge, lack of, 139 Knowledge translation (KT)
- economic evaluation, 170–173
- end-of-grant, 129–130
- identifying barriers to, 130–131
- integrated, 129
- interventions for, 131–132
- key questions on, 130
- open access publications and, 190–191
- planning for, 130
- presentation at scientific meetings, 191
- systematic review, 128–132
- tips for conducting, 133
Knowledge translation science, 129 Kruskal–Wallis test, 230t, 239–241

L
Language bias, 93t Leadership, 221–223, 222t Lead time bias, 50, 50f Least squares estimate, 247 Length time bias, 50, 51f Likelihood
- maximum, 276, 279–280
- restricted maximum, 276 Likelihood ratios, 35–39
- example of, 39–40, 39t
- in meta-analysis, 271–273
- on nomograms, 36–38, 37f
- positive, for multilevel outcomes, 38–39
Linear mixed effect models, 63 Linear regression
- diagnostics for, 252, 253f
- inference in, 248
- multiple, 249–252, 251f
- simple, 245–249, 246f
Literature search, 1, 119, 121f, 176–178, 183

M
Magnetic resonance imaging (MRI)
- of cervical cancer, decision analysis of, 149–150
- economic evaluation of, 153–154
- of fatty tumors, sensitivity of, 47–48, 47f
- of hemophilic arthropathy, 54–55, 76, 76f–78f, 77
- of hippocampus, reliability of, 63–64, 63f, 64f, 65f
- of lung tumor volume, responsiveness of, 77–78
- of rabbit knees, responsiveness of, 78, 79f, 80, 82, 82f
Mammogram
- film vs. digital, 43–44, 44f
- meta-analysis of, 270–271, 271t, 277–280
Mann–Whitney test, 230t, 238–239
MANTIS database, 119
Index detection of, 274–276, 275f
estimating, 276
ignoring, 276
incorporating, 276
quantifying, 276
testing for, 274–276
understanding/explaining, 276–277
of intervention studies, 270–271, 271f
random effects model for, 273–274, 279–280, 279–282
subgroup, 276–277
synthesis objective of, 269
systematic review for, 116–118 (See also Systematic reviews)
Meta-regression, 276–277
Methodological quality, definition of, 124
Misclassification bias, 93t
Mode, 21, 21f
Model fitting, 95
Modes, histogram, 15–17, 16f, 17f
Monte Carlo simulation, 169
Multimodal histogram, 17
Multiple adjustment, 96
Multiple linear regression, 249–252, 251f
Multiple logistic regression, 252–255
Multiple observations per participant, 202, 203f
N
Narrative reviews, 117, 118f
National Institutes of Health (NIH), 211, 213–214, 215f, 216f
Negative likelihood ratio, 35–36
Negative predictive value (NPV), 31–35, 39
Nested case-control study, 92
Noise, 58
Nominal data, 56, 229
Nomograms, 36–38, 37f
Nonexperimental research, 1, 2–9, 2f, 3f, 87–104
bias in, 92, 93t
comparative, 87
confounding in, 91, 92–98
economic evaluation in, 161
example of, 99–102
meta-analysis of, 269
practicality of, 87
reporting, 98
retrospective vs. prospective, 87–88, 198
STROBE statement of, 98–99, 100f
types of, 88
Nonparametric method, 44
Null hypothesis, 227–229, 229t
Numerical code, for data storage, 201
Nuremberg Code, 196, 197
0
Observational designs, 1, 2–9, 2f, 3f, 87–103
bias in, 92, 93r–94r
cumulative, 87
confounding in, 91, 92–98
economic evaluation in, 161
example of, 99–102
meta-analysis of, 269
practicality of, 87
reporting, 98
retrospective vs. prospective, 87–88, 198
STROBE statement of, 98–99, 100f
types of, 88
Observer bias, 61
Observer variability, 59, 60f, 61
Odds of the outcome, 254
Odds ratio (OR), 98, 123, 230t, 232–233
diagnostic, 271–274
in meta-analysis, 270–273, 277
sample size estimation for, 262–263
One-tailed test, in sample size estimation, 261
Ontario Health Insurance (OHIP), economic
data from, 161
Open access publications, 190–191
Opinion leaders, 131
Options, in decision analysis, 146, 149
Ordinal data, 56–57, 56f
Outcome Measures of Rheumatoid Arthritis
Clinical Trials (OMERACT), 56
Outcomes
in case-control studies, 8–9, 8f
clinical, meta-analysis of, 270–271, 271t
in decision analysis, 147, 150
in economic evaluation, 165–166, 165t
in observational studies, 98
in randomized controlled trials, 110, 114
in sample size estimation
continuous, 261–262
count, 261
proportion, 262–265
in systematic reviews, 123
Outcome variable, in linear regression, 245,
246f
Outliers, 22
Out-of-pocket costs, 158–159, 159t
Overmatching, 96–97
Overview Quality Assessment Questionnaire
(OQAQ), 127
Parameters, in decision analysis, 146–147, 149
Parkinson’s disease, ultrasound in, 49
Participation bias, 93t
Patient costs, 158–160, 159t, 160t
Paule–Mandel method, for estimating
heterogeneity, 276
Pearson’s chi-squared test, 230, 235
Pearson’s correlation coefficient, 68, 248–249
PeDRO database, 119
Peer review, 174, 192
Percentiles, 22
Performance bias, 93t, 122–123
Personal health information (PHI), 198–199,
201–202
Perspective
in decision analysis, 144, 146, 149
in economic evaluation, 163–165, 164f
PET/CT
of cervical cancer, decision analysis of,
149–150
of Hodgkin lymphoma, economic
evaluation of, 161
Philips checklist, 125
PICO approach, 1, 2, 176–177
Point estimation, 225–226
Pooled analysis, 117
Population, in statistical model, 225
Positive likelihood ratio, 35–36
Positive predictive value (PPV), 31–35, 39
Predictive validity, 73–74, 74f
Predictive value, 31–35, 39
Predictor variable, 245, 246f
Preferred Reporting Items for Systematic
Reviews and Meta-Analyses (PRISMA),
127, 182
Pretest probability, 36–38, 37f
Price, in economic evaluation, 160–162
Primary data, 1, 2f
Primary literature, 117
Printed educational materials, 131
PRISMA, 127, 182
Probabilistic sensitivity analysis, 169, 170f
Probability. See also specific statistical methods
in decision analysis, 138–142, 146–147, 149
in statistical model, 225–227, 227f
Productivity (indirect) costs, 158, 159–160, 160t
Prognosis, research design for, 10
Promotion requirements, in academia, 215–218
Proportion outcomes, sample size estimation for, 262–265
Prospective cohort design, 6–7, 6f, 11t, 89–91, 90f
Prospective studies, 87–88, 198
Protected time, for research, 219
Protocol phase, of systematic review, 119, 120–121t
PsycINFO, 119
Publication bias, 94t, 105, 127, 128, 191
Publishing, 174–195
  authorship and contributorship in, 178–179
  frequently asked questions, 192
  likelihood of, tips to improve, 193–194
  manuscript preparation for, 179–190
  manuscript submission for, 191–192
  open access, 190–191
  peer review in, 174, 192
  reasons for, 174
  review of galley proofs, 192
  trainee as primary author, 174–175
“Publish or Perish,” 174
Q
  Q statistic, 274–276
  Q test, 274–276, 278, 280
Qualitative data, 13
Quality-adjusted life-year (QALY), 147, 156, 166, 167, 170–171
Quality Appraisal of Reliability Studies (QAREL), 124
Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), 124, 124f
Quality control, data, 204
Quality of Health Economics Studies (QHES), 125
Quality of life, 54, 55, 71, 166
Quality of Life (QLQ)-C30 questionnaire, 83
Quality of Reporting of Meta-Analyses (QUOROM), 127
Quantile-based spread, 26
Quantile–quantile (QQ) plot, 252, 253f
Quantiles, measures of, 21–24
Quantitative data, 13
Quartiles, 22–24
Question, research. See Research question
R
  Radiographics, 88
Radiological Society of North America (RSNA)
Creating and Optimizing the Research Enterprise Workshop (CORE), 222
Educational Scholar Grant, 216t
Research & Education Fund, 214
Research Scholar Grant, 216t
Research Seed Grant, 216t
workshops on grant writing, 215t
Random effects (RE) model, 273–274, 279–282
Random error, 59, 60f
Randomization, 96t, 107–109, 112–114
Randomization sequence, 107–108, 112–114
Randomized controlled trials (RCTs), 1, 9–10, 11t, 88, 104–115
  allocation in, 9, 106, 107–108, 112–114
  allocation concealment in, 108–109
  analysis of, 110–112, 114
  bias and confounding in, 9, 10, 104–106, 108
  blinding in, 61–62, 105, 109, 114
  concepts of, 104–106
  CONSORT statement on, 112–114, 113f, 182
  cross-over in, 110, 269
  dropouts in, 110
  economic evaluation in, 160–161
  eligibility for, 9, 107
  equipoise in, 110
  example of, 112–114
  intention-to-treat analysis of, 110–112, 111f
  intervention in, 109–110, 114
  meta-analysis of, 269, 277–280
  outcomes in, 110, 114
  overview of, 106–110, 107f
  reporting, 114
  study sample for, 107, 112
Rate ratio (RR), 270–271
Ratio data, 57, 58f. See also Continuous data
Rationality, 146
Ratio-of-means (RoM), 270–271
RCTs. See Randomized controlled trials
Reader-order bias, 94t
Recall bias, 93t
Receiver operating characteristic curve (ROC), 40–43, 42f, 44f, 83. See also Area under the curve
Rectal cancer, data types on, 58, 59f
References, in manuscript, 185
Reference standard, 3, 30
  in case-control design, 8
  in cohort design, 6–7, 6f, 10
  in cross-sectional design, 5, 5f, 10
  imperfect, 51
Referral bias, 51, 93t
Regression adjustment, 97–98
Regression analysis
  meta-regression, 276–277
  multiple linear regression, 249–252, 251f
  multiple logistic regression, 252–255
  responsiveness in, 83
  sample size estimation for, 266–267
  simple linear regression, 245–249, 246f
Regression coefficient, b, 83
Regression line, 245–247, 246f
Regression model diagnostics, 252, 253f
Rejection region, 228–229
Relative risk. See Risk ratio
Reliability, 3–4, 4f, 10, 54–56, 55f, 62–69
  association vs., 68
  case study of, 63–68, 63f, 64f
  definition of, 62
  external, 63
  internal, 62–63
  intraclass correlation coefficient of, 63, 64–68, 260, 264
  random error and, 59
  sample size calculations and, 259–260, 263–264, 266
Reminders, 131–132
Repeat measurement bias, 94t
Reporting. See also Publishing
economic evaluation, 170–173
observational research, 98
randomized controlled trials, 114
Representativeness heuristic, 141
Research. See also specific topics and methods
  clinical, fundamental goal of, 1, 196
  ethics in, 196–199
  imaging-related, topics in, 12
  peer review of, 174, 192
  success in academia via, 210–224
Research designs, 1–12
  analytic, 2–3, 3f
  analyzing and interpreting data from, 229–243, 230t
  choosing appropriate, 10–12, 11t
  descriptive, 2, 3f
  for economic evaluation, 160–161
  experimental, 1, 2–3, 2f, 3f, 9–10, 104–115
  meta-analysis and, 269
  nonexperimental (observational), 1, 2–9, 2f, 3f, 87–104
  sample size based on, 257, 258t
Research Ethics Board (REB), 197
Research projects
  budget for, 2, 204–207
  co-investigators for, 175
documentation (manuscript), 179–190, 180t–181t
  literature search for, 1, 176–178
  preparation for conducting, 174
  publishing results of, 174–195 (See also Publishing)
  search for, 175
  statistician for, 178
  steps in, 1–2
  supervisors for, 175
Research question
  in economic evaluation, 157, 158b
  FINER characteristics of, 2
  formulating, 1, 2–3, 176
  research designs to address, 10, 11r
  searchable, creation of, 176
Research tools, 2f
Residual/error, in regression analysis, 247, 248, 252
Residual sum of squares (RSS), 250
Response bias, 94t
Review phase, of systematic review, 119, 121f
Reviews, 116–118. See also Systematic reviews
  RevMan (Review Manager), 276, 279
Risk difference (RD), 230–233, 270–271, 277
Risk estimation, difficulties in, 139–140
Risk ratio (RR), 98, 230t, 232–233, 264–265
  in meta-analysis, 270–271, 277–280
ROC. See Receiver operating characteristic curve
“Rolling back” decision tree, 145, 149t
S
Salary awards/support, 213–214
Sample restriction (specification), 94, 95–96, 96t
Sample size, 1
  calculations of
    motivation for, 257
    quantities of interest in, 257, 257t
    factors influencing, 257, 258t
Sample size estimation, 257–268
  accuracy in, 257
  based on confidence intervals, 257–259, 258f
  target CI width for count outcome, 261
  target CI width for difference in proportion, 262
  target CI width for ICC, 264
  target CI width for kappa, 263–264
  target CI width for proportion outcome, 262, 264–265
  based on hypothesis test, 257, 259, 260f
  based on number of predictors of outcome, 266–267
  concepts in, 257–260
  for conservative estimate, 262
  for diagnostic test performance, 259–260, 265–266
  examples of, 261–267
  for odds ratio, 262–263
  power in, 257, 259, 260f
  comparison of two means, 261–262
  diagnostic test (AUC), 265–266
  one-tailed test for, 261
  target power for continuous outcome, 261–262
  target power for difference in proportions, 262–263
  for risk ratio, 264–265
  Sampling bias, 51, 93f, 104–105
  Sampling variability, in statistical model, 227, 227f
Scales
  for imaging modalities, 54–55
  for study quality, 119–122, 122f
  Scatterplot, 26–28, 27f, 64f, 65f, 66
  in economic evaluation, 169, 170f
  in simple linear regression, 245–246, 246f
  Scatterplot matrix, 27–28, 28f
  Scientific meetings, presentation at, 191
  Scientist track, 219
Scoring systems, 54–55
Searchable research questions, 176
Searchable vocabulary, 176
Searching (literature search), 1, 119, 121f, 176–178, 183
Secondary data, 1, 2f
Secondary literature, 117
Selection bias, 51, 93t, 104–105, 122–123, 127
Self-selection bias, 93t
Sensitivity, 31–35
  bias and, 49–51
  case studies of, 47–49
  comparison of two tests, 44–47
  in decision analysis, 147, 148, 147f
  in literature search, 176
in low-dose abdominal CT for appendicitis, 39–47, 39t, 41f, 42f, 44f
  meta-analysis of, 271–273, 280–282, 281f
  in MRI for fatty tumor, 47–48, 47f
  ROC curve for, 40–43, 42f, 44f
  in ultrasound for appendicitis, 32–35, 32t, 34t, 272–273, 272t, 280–282, 281f
  in ultrasound for Parkinson’s disease, 49
Sensitivity analysis, 147, 168–170, 170f
Short Form (SF)-36, 55, 166
Signal, 58
Sign test, 230t, 241–243
Simple adjustment, 96t
Simple linear regression, 245–249, 246f
Skewness
  boxplot, 22, 23f
  histogram, 17, 17f, 22, 23f
Spearman’s correlation coefficient, 68, 248–249
Specification, to control confounding, 94, 95–96, 96f
Specificity, 31–35
  bias and, 49–51
  case studies of, 47–49
  comparison of two tests, 44–47
  in decision analysis, 148, 147f
  in 111indium WBC scan for infection, 48–49, 48f
  in literature search, 176
  in low-dose abdominal CT for appendicitis, 39–47, 39t, 41f, 42, 44f
  in measurement tools, 54–55
  meta-analysis of, 271–273, 280–282, 281f
  ROC curve for, 40–43, 42f, 44f
  in ultrasound for appendicitis, 32–35, 32t, 34t, 272–273, 272t, 280–282, 281f
  in ultrasound for Parkinson’s disease, 49
Spleen ultrasound
  lymphoid nodules in, 75–76, 75f
  patterns in children, 69–71, 69f, 71t
Spread
  measures of, 25–26
  quantile-based, 26
  Standard deviation, 26
  Standard error, 79, 227
  of effect measures, 271
  in meta-analysis, 271, 273
  in one group with binary data, 230–231
  in one group with continuous data, 237
  in regression analysis, 247, 248, 252
  of risk difference, 232
  in sample size estimation, 258, 259, 262, 264
  in two groups with continuous data, 238
  Standardized effect size (SES), 78–80
  Standardized mean difference (SMD), 123, 270–271
Standardized response mean (SRM), 78–80
Standards for Reporting of Diagnostic Accuracy (STARD), 124, 182
“State of the Art” series, 117
State-transition models, 168
Statistical analysis, 2. See also specific methods
Statistical inference, 225–244, 248
Statistical model, 225–229
  estimating confidence interval for $p$ in, 226–227, 227f
  estimating $p$ in, 226, 227f
  hypothesis testing in, 227–229, 229t
  sampling variability in, 227, 227f
Statisticians, 178
Statistics
  descriptive, 13–29
  effect size, 78–80
Stratification, 95, 96f, 97, 108
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), 98–99, 100f
Structured reviews, 116–117. See also Systematic reviews
Study examination bias, 93t
Study identifier (study ID), 198, 201–202
Study participants, 4
Study sample, 107, 112
Subgroup meta-analysis, 276–277
Subject bias, 61
Subject headings, 119, 177, 190
Subject variability, 59, 60f, 61
Submission requirements, for manuscript, 190
Supervisors, 175
Supplemental material, in manuscript, 182t
Symmetric data, 17
Synthetic goal, 117
Systematic errors, 59–61, 60f
Systematic reviews, 1, 116–137
  bias in, 119–124, 122f, 127–128
  blinding in, 119–122
  Cochrane Collaboration, 125–126, 133
critical appraisal of, 127–128
data extraction in, 123
data synthesis in, 123–126
definition of, 117
extraction of data in, 123
goals of, 117, 118
interpretation of results, 126–127
knowledge translation from, 128–132
limitations of, 128
narrative reviews vs., 117, 118t
outcomes in, 123
PRISMA tool for, 127, 182
protocol phase of, 119, 120r–121t
quality of study in, 119–124, 122f, 127–128
review phase of, 119, 121f
selection of studies, 119
steps for conducting, 118–123
studies of diagnostic tests in, 123–124, 124f
tips for undertaking, 132–133
Users’ Guide to Medical Literature, 126

T
  Tables, in manuscript, 187
  Tabulation, 234–235
  Tails, of histogram, 17
  T1 bench to bedside, 128–129
  Terminal nodes, 144
  Testing section, of manuscript, 179, 180
  Test–retest reliability, 3
  Test-review bias, 94t
  Test runs, 200
  Test statistic, 228
  Text words, for searches, 177
  Therapy, research design for, 10
  Thresholds
    and heterogeneity, 274
    and rejection of null hypothesis, 228–229
  Time
    in academic careers, 219, 220–221
    in case-control studies, 5
    in cohort studies, 5–7
    in decision analysis, 147, 150
    in economic evaluation, 167
    in retrospective vs. prospective studies, 87–88, 198
  Time losses, in economic evaluation, 160, 162
  Time-to-event outcome, 270–271
  Title page, 190
  Tornado diagram, 169, 169f
  Total variance, 274–276
  Trainee
    as primary author, 174–175
    search for supervisor, 175
  Transfer bias, 93t
  Transfer of copyright agreement, 182
  t-test, 237–243
    paired, 78–80, 230t, 241–243, 282
    power, in sample size estimation, 261–262
    single-sample, 230t, 237–238
    two-sample, 238
    unpaired, 230t
  Tversky, Amos, 141
  Type 1 error, 229, 229t
  Type 2 error, 229, 229t
Index

U
Ultrasound
for appendicitis
  bias in, 51
  economic evaluation of, 168–169
  meta-analysis of, 272–273, 272t, 280–282, 281f
  predictive criterion validity of, 74, 74f
  sensitivity and specificity of, 32–35, 32t, 34t, 272–273, 272t, 280–282, 281f
  systematic review on, 119, 120t–121t
Doppler coronary artery, responsiveness of, 82–83
for liver malignancies, statistical model of, 225–229
for spleen
  lymphoid nodules in, 75–76, 75f
  patterns in children, 69–71, 69f, 71t
Unbiasedness, 26
Uncertainty
  in decision making, 140–141
  definition of, 140
  in economic evaluation, 168–170
Uncertainty analysis, 147
Unimodal histogram, 17, 17f
Unstructured reviews, 117–118
Users’ Guide to Medical Literature, 126
Utilization quantity or volume, 160

V
Validity, 3–4, 10, 54–56, 55f, 71–77
  concurrent, 73, 75–76, 75f, 76f
  construct, 76–77, 77f, 78f
  content, 72–73, 74f
  convergent, 76–77, 77f
criterion, 73–76, 74f–76f
discriminant (divergent), 76, 77, 78f
face, 72, 72f, 73f
predictive, 73–74, 74f
  strategies to increase, 71
  systematic error and, 59–61, 60f
  tests of, 71–72
Valuation (cost), 158, 162–163, 163t, 164t
Values, in decision analysis, 142, 150
Variability, measurement, 59–61, 60f, 61
Variables. See also specific studies and techniques
  association or relationship between
decision tree on, 251f
  measures of, 26–28
categorical, 250–252, 251f
dependent, 245
dummy coding, 250–252
  independent or predictor, 245, 246f
Variance, 26, 260
Variance, total, 274–276
Verification bias, 50–51, 93t, 105
Visual aids, in manuscript, 182t, 187–188, 187f

W
Weighted kappa, 70–71
Weighted means, 18–19
Whiskers, of boxplot, 22–24
Wilcoxon rank sum test, 238–241, 260
Wilcoxon signed rank test, 230t, 241–243
Workup bias, 51, 93t, 105

Y
York Centre for Reviews and Dissemination, 171