



UASLP
Universidad Autónoma
de San Luis Potosí

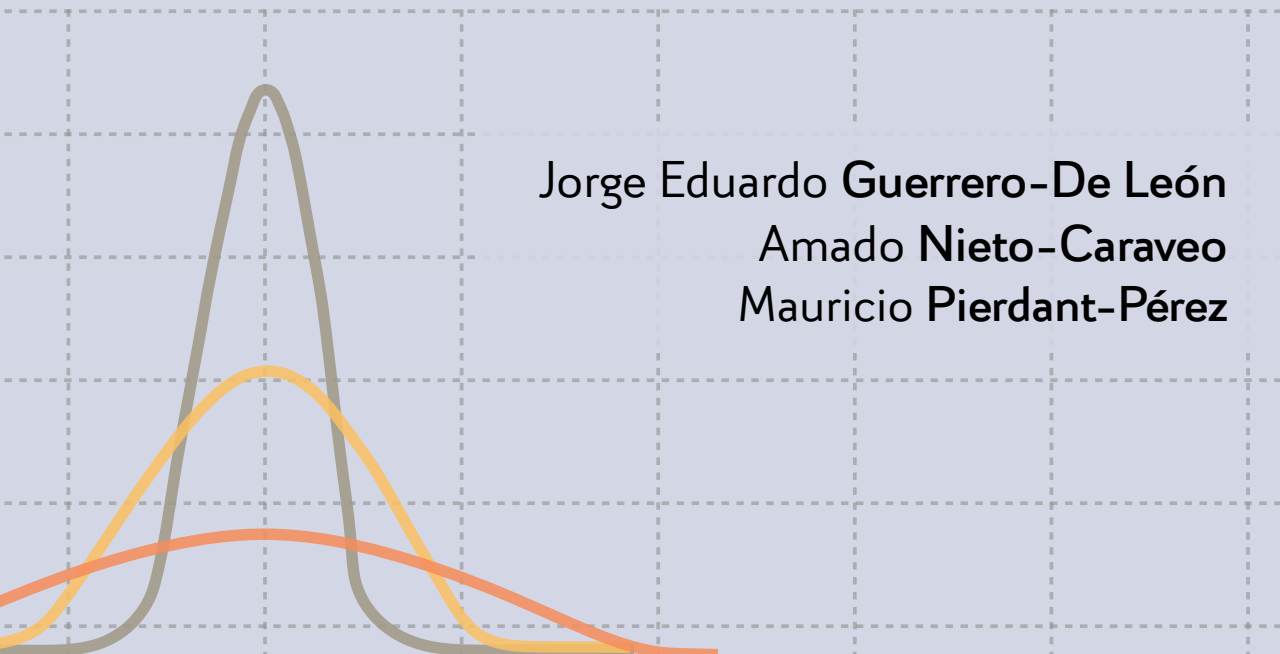


FACULTAD DE
MEDICINA



CLINICAL EPIDEMIOLOGY

An Active Learning Approach



Jorge Eduardo Guerrero-De León
Amado Nieto-Caraveo
Mauricio Pierdant-Pérez



CLINICAL EPIDEMIOLOGY

An Active Learning Approach

Editors

Jorge Eduardo Guerrero-De León, M.D.

Specialization in Education Development:
Active Learning and Digital Pedagogy;
Certificate in Cognitive Neuroscience
Facultad de Medicina

Universidad Autónoma de San Luis Potosí, México

Amado Nieto-Caraveo, M.D., MSc

Head of the Public Health and Medical Sciences Department
Facultad de Medicina

Universidad Autónoma de San Luis Potosí, México

Mauricio Pierdant-Pérez, M.D., MSc

Coordinator of the Medical Informatics Area
Facultad de Medicina

Universidad Autónoma de San Luis Potosí, México

Alejandro Javier Zermeño-Guerra, M.D.
Dean of the Universidad Autónoma de San Luis Potosí

Marco Antonio Aranda-Martínez, LL.M.
Secretary-General of the Universidad Autónoma de San Luis Potosí

Ismael Francisco Herrera-Benavente, M.D.
Head of the Facultad de Medicina

Jorge Eduardo Guerrero-De León, M.D.
Amado Nieto-Caraveo, M.D., MSc
Mauricio Pierdant-Pérez, M.D., MSc
Editors

Jorge Eduardo Guerrero-De León, M.D.
Design

Talleres Gráficos de la UASLP
Printing and binding

ISBN: 978-607-535-174-2

All rights reserved by

©Universidad Autónoma de San Luis Potosí
San Luis Potosí, October 2020

Universidad Autónoma de San Luis Potosí
Álvaro Obregón 64, Zona Centro, C.P. 78000,
San Luis Potosí, S.L.P., México.

Please notify us of any inaccuracies or errors in the book by sending an email to clinical.epidemiology@med.uaslp.mx.

To my loving mother and sister
in acknowledgment of
their unconditional support.

Jorge Eduardo Guerrero-De León

To my two daughters and four sons.

Amado Nieto-Caraveo

To my students and fellow professors of the
Clinical Epidemiology Department
for all their teachings and help in finding our path.

Mauricio Pierdant-Pérez

Contents

Foreword ix

Preface xi

Acknowledgments xv

Section I

Introduction to Clinical Epidemiology and Active Learning

Chapter 1. Introduction 03

Chapter 2. Evidence-Based Medicine 09

Section II

Searching and Appraising the Best Clinical Information

Chapter 3. Framing the Research Question 15

Chapter 4. Planning a Search Strategy 23

Chapter 5. Search Strategies 33

Section III

Basic Principles of Biostatistics

Chapter 6. Answering the right question: Why do I have to learn Biostatistics? 39

Chapter 7. Basic Principles of Biostatistics 41

Chapter 8. Getting Down and Dirty with Data 43

Chapter 9. Descriptive Statistics 55

Chapter 10. Inferential Statistics 77

Chapter 11. Statistical Tests 89

Chapter 12. Correlation and Regression 103

Section IV

Risk and Prognosis of Diseases

Chapter 13. Study Designs 117

Chapter 14. Cohort Studies 123

Chapter 15. Case-Control Studies 129

Chapter 16. Cross-Sectional Studies 135

Chapter 17. Survival Analysis 141

Chapter 18. Disease Occurrence, Risk, Association, Importance, and Implication 151

Chapter 19. Odds Ratio and Relative Risk: As Simple as It Can Get 157

Chapter 20. Confounding 167

Chapter 21. Attributable Risk 175

Section V

Use of Diagnostic Tests

- Chapter 22.** Likelihood Ratios **183**
- Chapter 23.** Pre and Post-Test Probability **193**
- Chapter 24.** Sensitivity, Specificity, and Predictive Values **201**
- Chapter 25.** ROC Curves **211**

Section VI

Clinical Interventions

- Chapter 26.** Introduction to Experimental Study Designs **219**
- Chapter 27.** Ethics in Clinical Trials **221**
- Chapter 28.** Randomized Clinical Trials (RCT) **227**
- Chapter 29.** Equivalence, and Non-Inferiority Trials **247**
- Chapter 30.** Efficacy, Effectiveness, Efficiency **255**
- Chapter 31.** Bias **261**

Section VII

Evidence Synthesis

- Chapter 32.** The Role of Evidence Synthesis in Health Care **273**
- Chapter 33.** Systematic Reviews and Meta-Analysis **277**
- Chapter 34.** Clinical Practice Guidelines **287**
- Chapter 35.** Quality of Evidence **299**

Appendices

- Appendix A.** Answers to Active Learning Multiple Choice Questions **311**
- Appendix B.** Bradford Hill's Criteria for Causality **313**

Glossary **319**

Index **331**

Foreword

For over 10 years, a group of professors at the Facultad de Medicina of the Universidad Autónoma de San Luis Potosí have taken up the challenge of revolutionizing the learning approach to Clinical Epidemiology. On this journey, we have learned a lot, and we have made a lot of mistakes, but we believe we have found a way (a small path) that has led us to where we are right now: we incorporated critical thinking skills and competencies in the decision-making process, we scrutinize the conventional schemes of evidence-based medicine and traditional clinimetry up to adopting a blended model with an active learning approach by the student. Our intention is to foster the competencies in our med students that will permit them to build up **clinical expertise** and continue their path through the transformation of their skills into a model of **adaptive expertise** with the years.

To become an adaptive medical decision-maker, students must create critical thinking skills and meta-cognitive processes such as reflection and mindfulness, which have to be mentored. Students need to construct the ability to transfer concepts learned in one specific context into new contexts and into novel situations. In this trail to **expertise**, our students necessarily begin in the analytic mode (System 2 from the two modes of thinking), because the recognition of patterns of symptoms and signs is not yet possible. At this point, System 1 from the two modes of thinking, the faster “intuitive” mode, is minimal. Through repetition and learning, students become more experienced to the point where basic patterns become familiar and will elicit System 1-based responses.

In order to firmly build their adaptive decision-maker abilities, first students need to establish an explicit acquisition phase, where processes become embedded in their cognitive and behavioral repertoires through learning (often over-learning). In this phase the templates for illness scripts begin to appear. It is in this path, if our students have an unquestioning, passive attitude, using learning by memory with minimal insight, where they will accumulate **experience**, but may not gain **expertise** (i.e., they can become experienced non-experts).

However, if they do possess insight, and actively engage with the clinical setting, they may progress instead toward proficiency and competence, showing efficient and accurate mastery of concepts to achieve “classic” or “routine” expertise. If we prepare them for this pattern, they will continue with their independent practice, and they will maintain a learning approach that develops qualities of the adaptive learning described above.

This will lead to the accumulation of experience with the varied presentations of a disease, and together with the capability of adopting innovative approaches toward novel and atypical cases, may progress beyond routine expertise towards **adaptive expertise**.

This book, brightly led by a young physician, Jorge Eduardo Guerrero-De León, shows the way to accomplish all we offer the students, but with a special approach: based on active learning. The student who wants to harvest medical decision skills in his future clinical practice must make his own attempt to fulfill it, and this book will serve as a bridge to facilitate him to carry it out.

Mauricio Pierdant-Pérez, M.D., MSc

Preface

“Like everything great that has been achieved in this world, this project began with an idea.”

This is how the Preface of my first book begins, a compendium of questions and answers about Histology that I made in my first year of professional career, and that in 2015 I dedicated to all new students of the MD program at the Facultad de Medicina of the Universidad Autónoma de San Luis Potosí, my alma mater. Five years later, I have reached the end of this adventure-filled roller coaster and feel that there is no better introductory line for the book you are holding.

Since 1936, Social Service has been a fundamental requirement to become a physician in Mexico. During this period, the student must integrate and apply the skills and knowledge acquired in their years of training in order to provide solutions to health-disease issues. It is in this context that, after an abrupt and unforeseen change that my life took, I was offered the opportunity to carry out my Social Service under a different modality than the one I aimed. I would conclude the last year of my MD program at my School, in a Medical Education-based program.

Education and teaching have been a significant part of my daily life over the last eight years. In all that time “Be who you needed when you were younger” has been the mantra that has allowed me to develop the passion I have for this disciplines.

Medical Education seeks to enhance the teaching-learning process of generations of physicians based on the principle that health is one of the most treasured public goods, since it allows us to pursue greater standards of well-being and progress. Training of human resources still involves providing sufficient tools with which the physician can acknowledge the demands and conditions that he will encounter throughout his professional performance within society.

The origins of Epidemiology go back to the initial studies carried out from outbreaks of infectious diseases or epidemics, conditions where the foundations of the research methodology were established. Currently, its influence has reached the clinical area, where epidemiological and statistical concepts are applied to establish research methodologies. This gives the trained physician in this discipline the ability to carry out clinical research, critically appraise studies published in the medical literature, assess the quality of health care, carry out economic evaluations, and many more.

Today, the medical information available is particularly extensive, often badly organized, and with high variability in the quality of their studies. In fact, it is believed that 85% of research resources are wasted. In recent years, different authors have highlighted the need to better assess medical literature through a further detailed analysis of the scientific research. However, there is no institutional culture where critical and rigorous analysis of research reports is encouraged. This promotes that doctors in training, as well as those who study a postgraduate degree, and even those who already practice medicine have a feeble critical capacity to appraise the medical literature, making it complex to decide what information should be incorporated into daily practice.

Despite the vast resources that can be reached with the Internet, there is no concise, precise and useful material with which to acquire the basic knowledge that allows us to build up, over time, the ability to critically read scientific literature. This was the starting point in the creation of this textbook. Although I must point out that no textbook will entirely fit into established academic programs, and that there is no ideal textbook for teachers, students, or for every teaching-learning situation, these resources provide many advantages to their readers, one of the most extraordinary being the provision of a structure on which learning is built.

One of the features that can be questioned about this book is why it is written in English. Debates have arisen in non-English speaking countries over the chosen language of instruction in Medical Education, whether it has to be the English language or the mother tongue. As globalization advances, more people become bilingual or multilingual, and researchers have clearly demonstrated that bilingualism leads to cognitive advantages over lifespan. When compared to monolinguals, bilinguals are faster in information processing and conflict resolution in nonverbal tasks. Furthermore, bilingualism induces experience-related neuroplastic changes in several brain areas such as the frontal lobes, the left inferior parietal lobule, the anterior cingulate cortex, and in subcortical structures such as the left caudate and putamen. These areas are part of the executive control network and may explain why bilinguals usually have a cognitive advantage in executive control tasks over monolinguals.

Physicians and researchers need to learn English, not only because it is the official language of the largest proportion of scientific literature related to Medicine in all its disciplines. Literature written in English contains the most up-to-date, current information that governs the best medical practices worldwide. English is a medium for teaching and learning, but is also the way to increase the visibility and the international diffusion of the research work that is done in our country.

Alcina-Caudet stated that “Despite the desire of some researchers to preserve the Spanish language as a language for the dissemination of their knowledge, it is sometimes impossible to stop this unstoppable inertia that leads to the use of a vehicular language other than the mother tongue. Even Ramón y Cajal himself had to give up his efforts to publish the magazine of his Institute in Spanish and went on to publish it in French, despite his well-known passionate defense of the Spanish language.”

The content of the 35 chapters and two appendices that make up this book is based on the instructional design of the Clinical Epidemiology course taught to the students of the MD program at the Facultad de Medicina of the Universidad Autónoma de San Luis Potosí. It is divided into seven sections that will allow the reader to acquire the basic tools to improve their ability to make medical decisions based on the appropriate formulation of a clinical question, carry out an adequate systematic search and retrieval of information, and to interpret clinical estimators for diagnosis, risk, prognosis and treatment of diseases in different clinical settings.

I hope and desire for this book to provide a practical, understandable and useful resource of information that will lead its readers to make appropriate clinical decisions based on the best available evidence, and consequently firmly impact on the quality of care of the patients, and in the health of our population.

I cannot finish this Preface without pointing out that nothing would have been achieved without the support, trust and opportunities that my tutor Amado Nieto-Caraveo, M.D., MSc, gave me, with whom I am and will ever be infinitely thankful.

Jorge Eduardo Guerrero-De León, M.D.

Acknowledgments

We are extremely grateful with the authorities of the Facultad de Medicina of the Universidad Autónoma de San Luis Potosí, especially with its current directives:

Jorge Luis García-Ramírez, M.D.; Scholar Services Secretary of the Facultad de Medicina.

Luis Guillermo Gerling-De Alba, Academic Secretary of the Facultad de Medicina.

Maribel Martínez-Díaz, M.D.; General Secretary of the Facultad de Medicina.

Ismael Francisco Herrera-Benavente, M.D.; Head of the Facultad de Medicina.

We also convey our sincere gratitude to the Dean of the Universidad Autónoma de San Luis Potosí:

Alejandro Javier Zermeño-Guerra, M.D.

For their trust and support by providing us the means so that “Clinical Epidemiology: An Active Learning Approach” would be delivered in a such a timely and profesional way.

Section I

Introduction to Clinical Epidemiology and Active Learning

Chapters of the Section

Chapter 1 Introduction

Chapter 2 Evidence-Based Medicine

Introduction

Learning objectives for this chapter

- A. Understand the importance of studying clinical epidemiology.
- B. Understand the importance of critically appraise clinical evidence.
- C. Learn about active learning as a tool to improve your medical education.

Here is your first “test” question:

Which of the following is your primary goal for this Clinical Epidemiology course?

- a) To learn most of the things related with Clinical Epidemiology.
- b) To spend as little time as possible studying Clinical Epidemiology.
- c) To get a good grade in Clinical Epidemiology.
- d) All of the above.

If you answered **A**, you have the ideal motive for studying Clinical Epidemiology—and any other course for which you have the same goal. Nevertheless, **this is not the best answer**.

If you answered **B**, here’s a simple suggestion: Drop the course, and your mission is accomplished.

If you answered **C**, you have acknowledged the greatest short-term motivator of many students in every School.

If you answered **D**, you have chosen the **best answer**.

There is nothing wrong in striving for a good grade in any course, just as long as **it is not your major aim**. Getting a good grade helps the grade point average, and all the consequences that this involves (such as getting a better place to choose where to do your internship practice).

But speaking about Clinical Epidemiology, this can lead to serious problems when you graduate, practice medicine, and do not have the skills to **critically appraise the medical literature**, leading to decisions not entirely based on the strongest evidence available.

There is also nothing wrong with spending as little time as possible studying this subject as long as you **learn the needed amount of stuff in the time spent**. The amount of time required to learn any subject depends on how you study and learn. Reducing the time required to complete any task satisfactorily is a worthy objective. It even has a name: **efficiency**.

Finally, there is nothing wrong with learning most of the things related with Clinical Epidemiology, as long as it **doesn't interfere with the rest of your schoolwork**, your **hospital practices**, and the **rest of your life**. Maintain some balance.

To summarize, the best goal for the Clinical Epidemiology course—and for all courses—is to **learn as much as you can in the smallest, but reasonable amount of time**.

“Clinical Epidemiology: An Active Learning Approach” is a book designed to provide you with many tools that will be helpful during the rest of your professional preparation.

If you think Clinical Epidemiology is a tough subject, **ask yourself and respond honestly**: “Is it really a complex subject or is it challenging for me because I have not established an ideal learning method?”

Your honest answer will tell you what to do.

Learning

Throughout life, human beings experience the learning process in many forms. In this experience, a series of internal and external factors occur simultaneously that can speed up or hinder the process.

Learning is a generic term for a diverse number of different cognitive processes. Its simplest and broadest definition can be encapsulated as: learning is a **change** in the state of a system produced by experience and **reflected in behavior**.

Learning is closely related to memory. Both terms must be distinguished from one another because sometimes they are misused as false synonyms. **Memory** refers to the states, conditions, images, or traces produced by the learning process that record what was learned.

Memory is often referred to as the “**engram**”, the physical, physiological, or neural change (embodiment of the stored information) that occurred when learning took place.

Learning comes in many different flavours, some of which appear to be the result of very simple neuronal changes and some of which appear to be inscrutably complex. Whatever they are, an enormous variety of learning types have been assayed over the years by researchers in this field, but they all collide in at least **nine different phases** intimately linked with each other:

- » **Motivation:** It triggers learning and is driven by the desire to learn, individual needs, and future prospects.
- » **Interest:** Expresses the student’s intention to achieve some objective and is intimately linked and conditioned by individual needs.
- » **Attention:** It is intimately linked with cognitive activities. Selective orientation of concentration and thought is the main phenomenon of attention.
- » **Knowledge acquisition:** It is the phase of the learning process in which the student initially gets in contact with the contents of a subject.
- » **Comprehension and internalization:** It involves thinking: the ability to abstract and understand concepts, as well as meaningful memory.
 - Comprehension is closely related to the critical capacity of the student. As content is understood, comprehension helps you judge it, relate it to previously acquired content and conceptualize the new cases presented.
- » **Assimilation:** The positive aspects of knowledge and experiences are stored and preserved in the medium and long term memory, thus affecting your subsequent behavior.
- » **Application:** The behavioral changes originated are usually strong when they are put into practice or “applied” in new situations, and have an effective and positive effect, originating a state of internal satisfaction.
 - If you do not apply assimilated knowledge, it generates frustration and its consequent loss.
- » **Transfer:** Meaningful learning has an effect on previously assimilated knowledge.
- » **Evaluation:** It allows observing and interpreting the results of the learning process and provides a starting point that will allow your process to be redirected, modified or maintained.

Performing these phases will favor the generation of knowledge and meaningful learning.

Active Learning

Conceptually, active learning is not an easy target. It is an umbrella term that embraces a variety of teaching and learning techniques, and represents a shift away from the exposition instruction that has a tendency to render learners bored or passive.

With active learning, students **take responsibility for their learning**, are **actively engaged** in building understanding of facts, ideas, and skills through the completion of instructor directed tasks and activities. This method emphasizes **higher-order thinking** and often involves **group work**.

Evidence shows that active learning is effective for maximizing learning, engagement, peer collaboration, and **evidence-based medicine**.

The goal of the active learning model is to train students to be proactive partners in the learning progression: to lean in and engage. Active learning can include different forms of activation, such as **interaction, social collaboration, deeper processing, elaboration, exploration of the material** and **meta-cognitive monitoring**. In addition, the various activities under the concept of active learning may involve different forms of instruction and are related to different cognitive processes. This is the basis for the design of this book.

The Textbook

Textbooks are one of the most important tools in most courses. Therefore, it is worth taking a few minutes to examine this book and look for its unique learning aids designed specifically to help you learn Clinical Epidemiology as efficiently as possible.

Learning objectives for each chapter

Each chapter begins with a list of goals that will tell you what you must be able to master after you finish studying the chapter. If you focus your attention on learning what is in these objectives, you will learn more in less time.

Sidebars

As you might have noticed, the side margins of the pages of this book are empty. They are intended for you to write in the margins, draw diagrams, highlight the key points...

In other words, this is your book, do whatever is necessary to help you fully understand and learn each chapter.

Cognitive bridges

Often in the study of any subject you see some term or concept that was introduced earlier in the course. To understand the idea in its new context, cognitive bridges help you to review it as it was presented earlier.

Add-ons

Sometimes it is necessary to give some extra information in order to get the full picture of a subject. That's why Add-ons are available thoroughly the chapters of this book.

Figures and Tables

Based on what we know about brain function and the vast number of neural connections between different brain regions, learning is facilitated when data are presented in multiple sensory modalities.

Visual elements such as graphs, charts, tables, and diagrams capture your attention, help to augment your written ideas, and simplify complicated textual descriptions. They will help you understand a complicated concept or visualize trends in the data.

End-of-chapter features

Immediately after the last section of each chapter, you will find the first set of this features: The **Key Terms List**. Key terms are essential to fully understand the subject you are learning. Define them as clear as possible. Key terms also appear in the Glossary, therefore we recommend you to use your Glossary regularly.

Following the Key Terms List is the **Active Learning Section**. Here you will find questions, exercises, and problems related to what you learned in the chapter. Some activities are relatively straightforward while others are more demanding. For the latter, we strongly advise you to **get together with some classmates** in order to fully engage in the study of Clinical Epidemiology.

- » As you solve the exercises in the book, remember that our main objective is for you to understand the principle upon which the exercise is based, not to get a correct answer.
- » Even when your answer is correct, stop and think about it for a moment. Don't leave the chapter until you feel confident with its contents.
- » Remember that in an Active Learning Approach, you are responsible for your learning, and must be actively engaged in building your knowledge. Try every available resource at your hands to thoroughly understand each subject until you become confident that you can solve any other problem in the near future.

Answers to multiple choice questions can be found in Appendix A. Finally, **Bibliography and Suggested Reading** is available in case you are feeling eager to learn more about the different subjects on this book.

Life is filled with moments when you must make decisions, specially in this career you chose to study. One fact you must realize is that **we live with the consequences that our decisions make.**

So choose intelligently, and **enjoy learning** whatever you set your mind to!

Bibliography and Suggested Reading

- Ausubel D. Adquisición y retención del conocimiento: una perspectiva cognitiva. Barcelona: Paidós; 2002.
- Bell D, Kahroff J. Active Learning Handbook. St. Louis, Missouri: Webster University; 2006.
- Geake J. Neuromythologies in education. Educational Research. 2008;50:2: 123-133.
- Goswami U. Neuroscience and education. British Journal of Educational Psychology. 2004;74: 1–14.
- McCoy L, Pettit RK, Kellar C, Morgan C. The goal of the learning-centric model is to train students to be proactive partners in the learning progression: to lean in and engage. Journal of Medical Education and Curricular Development. 2018;5: 1–9.
- Marzano R, Pickering D. Dimensiones del aprendizaje. México: ITESO; 2014.
- Navea-Martin A. El aprendizaje autorregulado en estudiantes de ciencias de la salud: recomendaciones de mejora de la práctica educativa. Educ Med. 2018;19(4):193-200.
- Prince M. Does Active Learning Work? A Review of the Research. J. Engr. Education. 2004;93(3):223-231.
- Yohannan DG, et al. Overcoming Barriers in a Traditional Medical Education System by the Stepwise, Evidence-Based Introduction of a Modern Learning Technology. Med.Sci.Educ. 2019;29:803–817.

Evidence-Based Medicine (EBM)

Learning objectives for this chapter

- A. Know the origin of the paradigm of Evidence-Based Medicine.
- B. Understand the advantages and limitations of this approach.

The main axis of the daily practice of health professionals is **decision-making**. The major purpose of these is to determine the what, when and how of interventions to be carried out, whether they are about diagnosis, treatment or prevention. The intricacy of the clinical environments where these decisions are made, has forced the building of different models that lead to the finest results for people's health. One of these models is that of **Evidence-Based Medicine (EBM)**, which has overcome over the past 30 years.

EBM is a process that intends to guide clinical decision-making based on a question asked about a specific problem, preferably "at the patient bedside".

Figure 2.1 shows the stages that this process follows. The "**information analysis**" component is frequently given higher emphasis, to the degree that many consider that EBM is limited to this component. We have given less importance to the adequate formulation of relevant clinical questions, to the systematic search and retrieval of information and to the application of this information in the clinical context of a specific patient. Without the clinical question, nothing that follows will make sense, no matter how sophisticated designs and analyzes are carried out. Similarly, information is meaningless by itself if it is not applied within a specific context in which decisions are made.

EBM arose in the **last decades of the 20th century**, in the conjunction of diverse movements originating in different places and with distinct motivations. **Alvan R. Feinstein's** ideas related to the application of statistical methods and the design of a "**clinical architecture**", created the basis of what was called the new "**Clinical Epidemiology**", a set of tools and methods that allowed for research rigorous a clinical hypothesis.

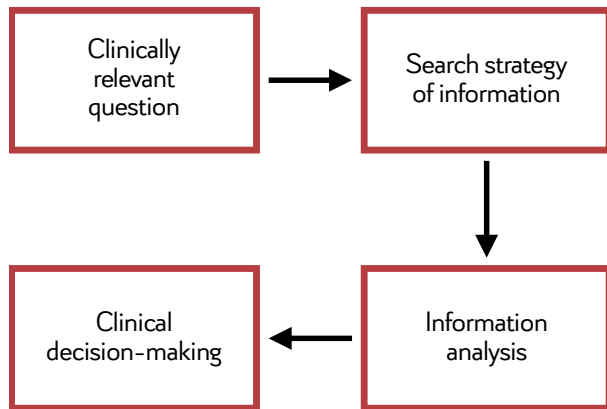


Figure 2.1. Stages of the Evidence-Based Medicine process.

In Great Britain, **Archibald Cochrane** promoted the conduct of controlled clinical trials on the assumption that most of the interventions that were carried out within the British health system were not only not effective, but also they meant a gigantic waste of resources. The great legacy of this initiative is the current **Cochrane Collaboration Project**, which reaches up to 400 systematic reviews per year.

The ultimate form of the EBM model was promoted at **McMaster University** in Canada by **David Sackett**, originally focused on the concept of “**hierarchy of evidence**”. Through this, it was possible to give value to the results of clinical research according to the type of epidemiological design from which they emerged. While this was a breakthrough, specially in the formulation of treatment guidelines, many have pointed out the limitations of this approach, which tend to downplay the results of well-conducted observational studies.

Another limitation of the EBM model is that it tends to ignore clinical reasoning mechanisms in specific contexts. The information from clinical research alone was expected to be sufficient for competent decision-making, but this has not occurred. On the contrary, in many situations the evidence can become so inconsistent or insufficient that it constitutes an obstacle to decision-making.

The **lack of reproducibility** observed in clinical research, the use of insufficient samples of patients, and the inadequate statistical analysis –just to mention some of the most common problems that have been reported–, generate a considerable number of uncertainty whose management is not generally considered by the EBM.

This has promoted the transformation of the classical model of EBM, towards more flexible forms that take into account the individual contexts where clinical decisions are made. Some have called this alternative model “**Evidence based ON medicine**”. Under this new perspective, the needs of patients are incorporated and the field of research is expanded not simply to the treatment and diagnosis of a disease, but also to the evaluation of health policies. There is also a need for better transparency and availability of the original clinical data.

EBM represents one of the **significant advances in medicine** in recent decades. Because of this, it has been possible to notify about the **overdiagnosis** and **overtreatment** of many common entities (e.g., prostate cancer and breast cancer), and it has also reversed conclusions that subsequently proved wrong (e.g., the use of Oseltamivir to prevent pneumonia in influenza infections).

However, the complexity in which health systems are currently established, along with the presence of multiple third parties involved, requires a model that allows, in the first place, to bridge the communication gap that exists between clinical research and decision-makers. To this extent, EBM will continue to be one of the privileged tools for improving people’s well-being.

Key Terms

Define the following terms.

Alvan R. Feinstein

Archibald Cochrane

Clinical architecture

Clinical epidemiology

Cochrane Collaboration Project

David Sackett

Decision-making

Evidence based on medicine

Evidence-based medicine

Hierarchy of evidence

Lack of reproducibility

Bibliography and Suggested Reading

- Barbara M. Sullivan, Ph.D. Essential EBP for Complementary and Alternative Medicine Study and Practice Guide. 2009.
- Nordenstrom J. Evidence-Based Medicine in Sherlock Holmes’ Footsteps. Stockholm: Blackwell Publishing Ltd; 2007.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: How to practice and teach EBM. 2° Edition. Edinburgh: Churchill Livingstone; 2000.
- Sheridan DJ, Julian DG. Achievements and Limitations of Evidence-Based Medicine. J Am Coll Cardiol. 2016 Jul 12;68(2):204-13.

Section II

Searching and Appraising the Best Clinical Information

Chapters of the Section

Chapter 3	Framing the Research Question
Chapter 4	Planning a Search Strategy
Chapter 5	Search Strategies

Framing the Research Question

Learning objectives for this chapter

- A. Understand where do clinical questions originate.
- B. Identify the different types of clinical questions.
- C. Learn the PICO system to formulate clinical questions.

Eugene Ionesco, father of the “Theater of the absurd” once said: “It is not the answer which enlightens, but the question.”

This absolutely applies to health care research because new knowledge originates from previously asked **answerable questions**.

In order to find new and useful answers to unresolved relevant problems, first we need to know a lot about the problem. Without this extense knowledge, it becomes difficult to imagine the development of plausible diagnostic tests or interventions. Further more, it becomes difficult to determine if we are “headed in the right ‘next-step’ direction”.

Researchable questions come from finding the “cutting edge” of knowledge for a health problem with which we are familiar.

In applied research, developing a question is an **iterative process**, not a “light bulb” phenomenon. That means that there is much work to be done before the light will shine.

The basic dimensions of a problem that drive to the formulation of important research questions include:

- » Understanding the Biology and Physiology of the problem.
- » Understanding its epidemiology (i.e., determinants and distribution, prevalence, incidence and prognosis).
- » Determining frustrations in the clinical management that have lead to unsatisfactory results for the patients.

Once these essential issues have been addressed, an initial direction for a question seems promising.

Where do Research Questions Originate?

Research questions can result from virtually **any point in the clinician's practice** with patients. Nevertheless, most of the questions derive from the following 6 aspects of clinical practice:

- 1. Clinical evidence:** How to gather clinical findings properly and interpret them accurately.
- 2. Diagnosis:** How to choose and interpret diagnostic tests.
- 3. Prognosis:** What to expect from the patient's likely course.
- 4. Therapy:** How to choose treatments that do more good than harm.
- 5. Prevention:** How to screen and reduce the risk for disease.
- 6. Education:** How to teach yourself, the patient, and the patient's family what is needed to be done.

This list can be kept handy and may be used as a “map” to where clinical questions come from.

Types of Clinical Questions

Clinical questions can be divided into two main categories: **background** and **foreground** questions.

Background clinical questions

- » Can be answered with information from textbooks, websites, and hospital or office resources, such as patient history files.
- » Focus on a **specific concept** (an intervention, an aspect of a disease or disorder, the determination of possible therapies).
- » Generally begin with a **question root** such as who, what, when, why, or how.

Foreground clinical questions

- » Seek to find relevant, sometimes individualized, evidence from **primary research publications**.
- » Typically include **component terms** or **keywords** focused on the patient, intervention, comparison to the intervention and outcome desired.
- » They are useful for **decision-making** in Medicine.

Another way to divide clinical questions is to classify them into **primary** or **secondary**.

Primary questions

What if too many questions arise?

For patients who have more than one active issue, and with possible questions about diagnosis, prognosis, and therapy for each problem, the questions may be too numerous to even ask, let alone answer. In this predicament, it's recommended to build good questions, then selecting the few questions that are most significant to answer.

A tip that can be applied in order to adequately choose the right question is to try this sequence of queries:

- » What is the most important issue for this patient now?
- » What issue should I address first?
- » Which question, when answered, will help me most?

The rest of the questions can be considered as **secondary questions**.

All primary questions must be asked “up front,” when the investigation begins. The same applies, as far as possible, for all secondary questions. This method ensures that the questions are “**hypothesis-driven**” (i.e., based on your predictions of what will happen) rather than “data-driven” [i.e., made up after the study results are (partly) in, especially to “explain” findings that may well be simply the play of chance].

This approach also allows for proper planning and data collection for these additional questions.

The PICO System

The formulation of a focused clinical question containing well-articulated elements is widely believed to be the **key to efficiently finding high-quality evidence**, and also the **key to evidence-based decisions**.

Following a structured method to formulate a clinical question must become a natural process in order to save time while questioning reality.

PICO frames were originally developed for therapy questions, but were later extended to all types of clinical questions. Using PICO frames has numerous **advantages**: it improves the specificity and conceptual clarity of clinical problems, elicits more information during the pre-search stage, leads to more complex research strategies, and yields more precise search results.

The PICO elements include:

- » **P**roblem/**P**atient/**P**opulation.
- » **I**ntervention/**I**ndicator.
- » **C**omparison.
- » **O**utcome.

In an optional fashion, a “**T**” can be added to the PICO acronym, forming **PICO(T)**. This represents **T**ime element or **T**ype of Study.

The components of a good clinical question can be thought of as data fields that will aid in the search for evidence and answers. The component terms can be used as key text words when using a search engine or **database management system** (DBMS – a “search engine” for databases). In addition, databases are often set up with searchable data fields, indexed terms and controlled vocabulary terms such as the National Library of Medicine’s PubMed **Medical Subject Headings** or “MeSH terms.”

Components of the Clinical Questions

Tailoring the clinical question’s component terms will help define and refine searches of medical literature databases.

Each component term can be used as a **search term** when searching the Internet, search engines or databases using a database management system (DBMS).

Components in background questions

Background questions are composed of fewer components, broader terms, and return more numerous (and sometimes less relevant) results.

Background terms can usually be replaced with specific synonyms found in the controlled vocabulary links of a database in order to produce a more specific search.

Components in foreground questions

Foreground questions typically use three or four component terms, use more specific terms, and often return fewer but more relevant results.

Framing the PICO-based question. Step-by-Step Tutorial

By following the PICO system, we're graphically and formally representing the mental process of an expert who asks questions.

Problem/Patient/Population

Defining the patient characteristics is **fundamental**.

A specific, narrow definition will provide truly applicable evidence for that particular patient, but may limit the evidence too much so that important evidence is excluded from the search results.

- » **Race** or **sex** can be essential to some health issues, but their inclusion **may limit** the retrieved results.
- » The search of information must be done **with and without** some **terms and limits** such as age, sex and race.
- » The key is to **be specific without becoming too narrow**.
- » Consider keywords and phrases that will allow a health care provider imagine the patient in front of you.
- » Try not to include extraneous information or terms.

Intervention/Indicator

This component may be broad or narrow.

When seeking “best evidence,” several interventions may be specified in separate foreground questions.

Broad phrases (“What is most effective?”) often lead to **background questions**.

- » Searching for background information from reliable, high-quality resources such as current textbooks, guidelines, reference handbooks and websites like Natural Standards, Natural Medicine, AHRQ, and MedLine Plus can help narrow the intervention component, so a good **foreground clinical question** can be created based on a background question.

A specific intervention should suggest something that will influence the desired outcome.

Comparison

This component is often considered the “second half” of the intervention component.

- » In **therapy questions**, intervention might be compared to a well-known or standard therapy.
- » For **diagnosis questions**, the comparison is often made to the “gold standard” diagnostic tool.

» **Prognosis or etiology** questions may include a factor which may affect the population in some way.

- Including symptoms (e.g., chronic cough) or exposure factors (e.g., second hand smoke) may provide a way to narrow the search without excluding essential results.

Outcome

The “outcome” is what the clinician **hopes to accomplish**.

An outcome should be **patient-oriented** (taking patient values, expectations, preferences and priorities into consideration), definable, measurable, and **clinically relevant**.

There may be cases in which there will be more than one relevant, important outcome. Each outcome can be defined in a separate PICO question.

Outcomes **should not be vague** (“feel better”) since vague phrases are not measurable and will not help to define a search strategy for significant evidence.

Outcomes such as “decrease pain”, “decrease the time to return to normal activities” and “increase physical function”, which can be defined and measured, may restrict the search, but should be considered when assessing results from a search.

In a few words, the outcome **is the most important component a PICO-based question**. It must be **as relevant as possible**.

Final considerations for a PICO question

Ninety-nine–word questions are difficult to comprehend, so it’s recommended not to get much detail in the question itself, but it is essential to bear these details in mind when conducting the study and reporting its results, so that they will not be overgeneralized.

Key Terms

Define the following terms.

Answerable clinical question

Background clinical question

Clinical evidence

Foreground clinical question

PICO System

Primary questions

Secondary questions

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Convert the following clinical cases into answerable clinical questions based on the PICO system and identify what sort of question it is.

- At a routine immunization visit, Lisa, the mother of a 6-month-old, tells you that her baby suffered a nasty local reaction after her previous immunization. Lisa is very concerned that the same thing may happen again this time. Recently, a colleague told you that needle length can affect local reactions to immunization in young children but can't remember the precise details.
- In browsing one of the medical weeklies, you come across mention of imiquimod cream for treatment of basal cell carcinomas (BCC). The idea of a cream for BCCs is surprising, so you wonder about the effectiveness and particularly the long-term cure rate of imiquimod cream.
- Susan is expecting her first baby in two months. She has been reading about the potential benefits and harms of giving newborn babies vitamin K injections. She is alarmed by reports that vitamin K injections in newborn babies may cause childhood leukemia. She asks you if this is true and, if so, what the risk for her baby will be.
- Mrs Smith has acute lower back pain. She has never had such pain before and is convinced that it must be caused by something really serious. You take a history and examine her but find no indicators of a more serious condition. You reassure her that the majority of acute low back pain is not serious but she is still not convinced.
- As part of your clinic's assessment of elderly patients, there is a check of hearing. Over a tea room discussion it turns out that some people simply ask and others use a tuning fork, but you claim that a simple whispered voice test is very accurate. Challenged to back this up with evidence, you promise to do a literature search before tomorrow's meeting.
- Childhood seizures are common and frightening for the parents and the decision to initiate prophylactic treatment after a first fit is a difficult one. To help parents make their decision, you need to explain the risk of further occurrences following a single seizure of unknown cause.

Bibliography and Suggested Reading

- Davies KS. Formulating the Evidence Based Practice Question: A Review of the Frameworks. *Evidence Based Library and Information Practice* 2011, 6.2:75-80.
- Formulate a Clinical Question. University of Western Australia. 2015. Available from: http://www.meddent.uwa.edu.au/teaching/acq/lo1acq_fcq/translate/pop_picoex.
- Haynes RB, Sackett DL, Guyatt GH, Tugwell P. *Clinical Epidemiology. How to Do Clinical Practice Research*. 3rd Edition. Philadelphia: Lippincott Williams & Wilkins; 2006.
- Huang X, Lin J, Demmer-Fushman D. Evaluation of PICO as a Knowledge Representation for Clinical Questions. 2006 Annual Symposium of the American Medical Informatics Association (AMIA 2006), November 2006, Washington, D.C.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: How to practice and teach EBM*. 2nd Edition. Edinburgh: Churchill Livingstone; 2000.
- Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making* 2007; 7:16.
- Sullivan BM. *Essential EBP for CAM Study and Practice Guide. Asking: Using the PICO Format to Structure a Search for Evidence*. 2009.

Planning a Search Strategy

Learning objectives for this chapter

- A. State the importance of an adequate search and retrieval of information to answer a clinical question.
- B. Understand Boolean Operators and its adequate use.

Once we have formulated a searchable question we need to build a **search strategy** in order to maximize the chances of finding a manageable number of relevant results.

First, let's begin by defining what a search strategy is.

Search Strategy

A search strategy is the **planned and structured organization of terms used to search a database**. The complexity of the strategy depends on the reasons for searching and the type of question you are investigating: If you are looking for background information, a simple search strategy may be enough. On the other hand, a literature review will require in-depth research with a comprehensive search strategy to ensure that all relevant sources of evidence have been identified.

A well-planned search strategy has enormous **benefits**, such as:

- » Retrieval of relevant references.
- » Inclusion of all key references.
- » Manageable number of results.
- » Efficient use of time.

Planning a search strategy includes the following **steps**:

- » Identifying search terms.
- » Truncation, wildcards, and phrases.
- » Combining terms with Boolean Operators (AND, OR, or NOT).
- » Applying limits to the search.

Identifying Search Terms: Concepts and Synonyms

Once the question has been formulated using the PICO method, the next step involves identifying the **words** that will be used to find information. These words are the **basis for the search**. You can start your search strategy by keeping a record of significant words while doing background reading on the subject, or by using a **thesaurus** or **dictionary**.

Concepts

Concepts are the **main topics or headings that emerge from the question**. It is suggested to keep to a maximum of four or five concepts to avoid over-complicating the search.

Synonyms

Synonyms are **alternative search terms** for the same concept. Terms can vary between countries, for example: “Accident and Emergency” in the UK has the same meaning as “Emergency Department” in the US.

Synonyms are included under broad subject headings in sophisticated databases such as MEDLINE, Embase, or PsycINFO. Word spellings may also vary: “paediatrics” in Australia is expressed as “pediatrics” in the USA.

Let’s consider the following PICO question: “Can brief intervention methods be used as an effective smoking cessation technique with teenagers?” We can identify the concepts and synonyms stated in **Table 4.1**.

Table 4.1. Concepts and synonyms from the example

	P	I	O
Concepts	Teenagers	Brief intervention	Smoking cessation
Synonyms	Adolescents Young people	Brief advice Brief counseling Motivational interview	Quit smoking Stop smoking

Keywords

Keywords are words that come to you naturally, or that may be **part of a specific discipline vocabulary** (e.g. terms used only by physicians), or that you brainstorm when planning your search.

If the “map term to subject heading” box is ticked in the database, it will attempt to map your keyword or phrase to a subject heading in the database’s thesaurus. If this option is not ticked, the word or phrase will be treated as a keyword.

Searching by keyword finds only those results where your keyword appears as an **exact match** in several fields including the title or abstract. This works particularly well if you are looking for a specific spelling, product, term, or phrase.

Remember: it is important to recognize that different spelling and terminology may exist for the same search topics. Keyword searching will not differentiate between spellings.

Subject Headings

Subject headings are **terms that have been identified and defined to cover a particular concept**. Synonyms are then “mapped” to those subject headings.

Databases like MEDLINE and Embase use a thesaurus to group related concepts together. In MEDLINE, the thesaurus is known as **Medical Subject Headings (MeSH)**; in Embase, the thesaurus is known as Emtree.

Table 4.2 shows a comparative between subject headings and keywords.

A search may include both subject heading and keyword features.

Broadening or Narrowing Search Terms

For each concept it is important to consider broader and narrower terms that may extend or limit the search. These may be useful for extending very specific terms, or limiting those where too many results are retrieved.

» An example might be a search that requires Mexican data. The location “Mexico” could be narrowed to a particular state if a large number of results are retrieved.

A search may be **broadened** by choosing a more general term and “exploding” to include all its associated sub-terms.

Related terms are those linked to the search terms by subject matter. For example, terms related to pregnancy might be prenatal care, pseudopregnancy, fetal, maternal-fetal relations, and pregnant women.

Table 4.2. Differences between Subject Headings and Keywords

Subject Headings	Keywords
Pre-defined “controlled vocabulary” words assigned to describe the content of each item in a database or catalogue	Natural language words describing the topic of interest
Databases look for subjects only in the subject heading or descriptor field, where the most relevant words appear	Databases look for keywords anywhere in the record (title, author name, subject headings, abstracts, etc.)
If a subject heading search yields too many results, you can often select subheadings to focus on one aspect of the broader subject	Often yields too many or too few results
Results are usually very relevant to the topic	May yield irrelevant results to the topic
Will find synonyms, plurals, spelling variations	Will not pick up synonyms, plurals, or spelling variations
Will locate most of the relevant papers, but in order to perform comprehensive search you may need to supplement your subject heading searches with some keyword searches	Necessary when there is no subject heading available for the concept you wish yo search

Boolean Operators

Boolean operators are specific words used to **combine** concepts or keywords in order to improve the chances of finding relevant information.

The most commonly used boolean operators are **AND**, **OR**, and **NOT**.

In order to clarify the use of this tools, let’s review the **Boolean Logic** with example created at the Ithaca College Library in the US.

Or

In database searching, “**OR**” **expands** a search by broadening the set. It is often used to combine synonyms or concepts.

» Strawberry **OR** chocolate **OR** vanilla = an ice cream lover with global tastes (**Figure 4.1A**).

A textword search for information about teenagers should include the words most commonly used: adolescents OR adolescence OR teens OR teenagers OR young adults.

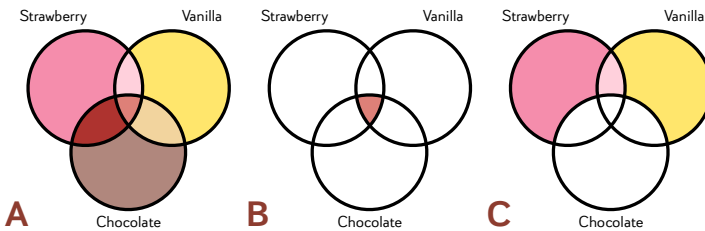


Figure 4.1. Graphic examples for the use of Boolean Operators

And

In database searching, “**AND**” **narrows** a search. It is often used for linking together different concepts.

» Strawberry **AND** chocolate **AND** vanilla = an ice cream lover who only eats ice cream which combines 3 flavours at the same time (i.e. Neapolitan ripple) (**Figure 4.1B**).

Not

In database searching, “**NOT**” is used to **get rid of** an unwanted concept.

» (Strawberry **OR** vanilla) **NOT** chocolate = an ice cream lover allergic to chocolate (**Figure 4.1C**).

Combining operators

You can combine sets in a variety of ways using combinations of Boolean Operators. When writing out the sets, **parentheses** are important because they define the order in which the concepts are processed.

» “**AND**” takes **priority** and is processed first - unless you use brackets to group concepts.

Suppose you have to find information on the incidence of salmonella food poisoning caused by hamburgers, with or without eggs. So, we could have the next scenarios:

» Salmonella **AND** hamburgers **AND** eggs = food poisoning caused by both.

» (Hamburgers **OR** eggs) **AND** salmonella = food poisoning caused by either.

» Hamburgers **OR** eggs **AND** salmonella = food poisoning caused by eggs, as well as hamburgers which have salmonella and which don't (the “**AND**” is processed first).

The statement, “students **AND** behavior” will only retrieve records in which the words ‘students’ and ‘behavior’ appear in the same document.

If you were interested in information on university students but not high school students, you could search (university students) **NOT** (high school). **BEWARE!** will exclude articles which discuss both types of students.

Bridge to Math
The use of parentheses is similar as brackets and parentheses are used in mathematical equations

» (Hamburgers AND salmonella) NOT eggs = only food poisoning caused by hamburgers, removing the confounding effect of having eggs at the same time.

The boolean operators are summarized in **Table 4.3**.

Applying Limits to the Search

It is important to determine what you are not searching for. Setting limits on the search is a way to further **refine** it to make your results more specific and relevant. Setting limits is an important part of your search strategy.

Common limits include:

- » Language.
- » Date of publication.
- » Type of publication (such as journal articles, book reviews, dissertations, reports etc).
- » Type of study (such as randomized controlled trial, cohort study, etc.).
- » Gender.
- » Age group.

Table 4.3. Summary of the Boolean Operators

Boolean operator	Purpose	Example	Result
AND	Combine keywords that reflect different concepts	falls AND aged	Each search result will contain both the terms falls and aged
OR	Combine keywords that reflect similar concepts	falls OR aged	Each search result will contain either (or both) the terms falls or aged
NOT	Exclude a keyword	(falls AND aged) NOT home	Each search result will contain both the terms falls and aged but only if they do not contain the third term home

- » Full text.
- » Abstract.

Refining Your Search

Once you have completed your search, you may find yourself in one of the following two scenarios: Too many or too few results were retrieved. Here's what you have to do:

If **too many results** were retrieved, go back over your strategy, and **NARROW** the search:

- » Use more specific terms as keywords.
- » Add terms for other aspects of the questions (e.g. age or gender of the patient), using the boolean operator AND.
- » Use more specific or relevant subject heading terms.
- » Use limits.

If **too few results** were retrieved, go back over your strategy, and **WIDEN** the search:

- » Use more terms: synonyms, related terms, broader terms.
- » Add terms with related meaning with the boolean operator OR.
- » Combine results of thesaurus and keywords.
- » Reduce or broaden limits (e.g. date range).
- » Select all subheadings of a subject heading term.

Key Terms

Define the following terms.

Boolean operators

Concepts

Database

Keywords

Limits to the search strategy

MeSH

Redefining the search strategy

Related terms

Search strategy

Subject headings

Synonyms

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Take a PICO-based clinical question (it can be one from the last chapter's Active–Learning Section or it can be a new one), and perform a search strategy in order to answer that question. You can take the following chart as a guide.

Question:

	Question part	Question term	Synonym
P	Problem/Patient/Population	() AND
I	Intervention/Indicator	() AND
C	Comparison	() AND
O	Outcome	() AND

Search terms used:

Hits

Results of the search strategy:

2. Report back on what you found out during your literature searching session. Discuss **WHAT** you found and **HOW** you found it. Try to include empirical evidence.
3. Use the Boolean Logic to describe women who have breast cancer but have never smoked and women who have breast cancer and who are EX-smokers

Bibliography and Suggested Reading

- Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making*. 1995;15:113-9.
- Haynes RB, Sackett DL, Guyatt GH, Tugwell P. *Clinical Epidemiology. How to Do Clinical Practice Research*. 3rd Edition. Philadelphia: Lippincott Williams & Wilkins; 2006.
- Henderson J & Strickland J. Boolean Logic. Ithaca College Library Research Guide. Accessed at <http://www.ithacalibrary.com/sp/subjects/boolean> on May 21, 2012.
- Henegan C, Badenoch D. *Evidence-based Medicine Toolkit*. 2nd Edition. Malden, Massachusetts: BMJ Books; 2002.
- Thurley N. & Coxall O. Finding the evidence 2 - Turning search terms into a search strategy. Bodleian Libraries, University of Oxford. Available from: https://www.youtube.com/watch?v=yh67_7zCQnA.
- University of Leeds Library. Developing your search strategy – Leeds University Library. University of Leeds Library. 2017. Available from: <https://library.leeds.ac.uk/researcher-literature-search-strategy>.
- Weinfeld JM, Finkelstein K. How to Answer Your Clinical Questions More Efficiently. Asking focused questions and knowing where to look can lead to quicker answers. *Fam Pract Manag*. 2005 Jul-Aug;12(7):37-41.

Search Strategies

Learning objectives for this chapter

- A. Acquire the basic tools to perform the search and retrieval of clinical information in a systematized form.
- B. Identify the main medical databases where you can obtain clinical information.

Search Strategies for Background Information

Remember that a **background question** asks for general knowledge about a disorder, disease, policy issue, etc. Consequently, background information may be found in sources such as:

- » Reference book entries or selected Ebooks/Encyclopedias in the health sciences.
- » Textbooks, chapters, appendices.
- » Drug monographs.
- » Guides to diagnostic tests.
- » Ebook drug guides.

Table 5.1 details some search strategies for background questions.

Search Strategies for Foreground Information

Remember that a **foreground question** seek evidence to answer a need for clinical information related to a specific patient, an intervention or therapy.

As such, identifying the PICO (T) elements helps to focus your question.

Table 5.2 details some search strategies for foreground questions.

Table 5.1. Search Strategies for Background Questions

Question	Appropriate source type	Sample strategy(ies)
What are the side effects of acetaminophen?	Drug reference book	Start with: StatRef (collection of reference tools, including drug references). Search on drug name. Access Medicine
What is Asperger Syndrome?	Textbook, monograph, review article	Start with: StatRef (collection of reference tools, including drug references). Search on name of disease or condition. Access Medicine Search Medline/ Pubmed for article type: "Review."
Evidence of the relationship between dementia and caffeine consumption.	Popular and scholarly article databases	Start with: Search Medline/ Pubmed for article type: "Review."
I need an overview of gestational diabetes	Textbook, monograph, review article	Start with: StatRef (collection of reference tools, including drug references). Search on name of disease or condition. Access Medicine UpToDate

Table 5.2. Search Strategies for Foreground Questions

Question	Natural language terms	Terms translated to Subject headings/ MeSH terms/ Descriptors [Database]
Does hand washing among healthcare workers reduce hospital acquired infections?	Hand washing Hospital acquired infections	Hand disinfection [MeSH] AND Cross infection [MeSH]
What is the effectiveness of continuous passive motion therapy (CPM therapy) following knee replacement in achieving optimal range of motion?	CPM therapy Knee replacement	Arthroplasty, replacement, knee [MeSH] AND Motion therapy, continuous passive [MeSH]
What is the effectiveness of restraints in reducing the occurrence of falls in patients 65 and over?	Falls Restraints	Accidental falls [MeSH, CINAHL] AND Restraint, physical [MeSH, CINAHL]
Does having access to fresh fruits and vegetables in neighborhood stores affect nutritional health of Hispanic Americans living in urban areas?	Food shopping, grocery shopping, grocery stores, food stores, bodegas, convenience stores Fruits, vegetables Hispanics Urban, city, cities	"Food Supply" [Mesh] AND ("Fruit"[Mesh] OR "Vegetables" [Mesh]) AND "Hispanic Americans" [Mesh]
I am looking for evidence-based articles on managing acute pain in sickle cell patients	Pain Sickle cell	In PubMed search: pain/therapy AND anemia, sickle cell then use Limits to limit to the Subset: Systematic Reviews or, Use Limits to limit to Article Type: Randomized Controlled Trial or Meta-analysis

Key Terms

Define the following terms.

Background clinical question

Foreground clinical question

Search strategy

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- 1. List some appropriate source types where you can find useful information to answer background clinical questions.**
- 2. List some appropriate source types where you can find useful information to answer foreground clinical questions.**

Bibliography and Suggested Reading

- Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical Epidemiology. How to Do Clinical Practice Research. 3rd Edition. Philadelphia: Lippincott Williams & Wilkins; 2006.
- Henegan C, Badenoch D. Evidence-based Medicine Toolkit. 2nd Edition. Malden, Massachusetts: BMJ Books; 2002.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: How to practice and teach EBM. 2nd Edition. Edinburgh: Churchill Livingstone; 2000.

Section III

Basic Principles of Biostatistics

Chapters of the Section

Chapter 6	Answering the right question: Why do I have to learn Biostatistics?
Chapter 7	Basic Principles of Biostatistics
Chapter 8	Getting Down and Dirty with Data
Chapter 9	Descriptive Statistics
Chapter 10	Inferential Statistics
Chapter 11	Statistical Tests
Chapter 12	Correlation and Regression

Answering the right question: Why do I have to learn Biostatistics?

Learning objectives for this chapter

- A. Understand the role of Biostatistics in modern medicine.

Biostatistics can provoke many medical students to feel frustrated and frantic, that's a fact.

So, why does the medical schools curricula includes Biostatistics?

Why does the Examen General para el Egreso de la Licenciatura en Medicina General (**EGEL-MEDI**) by CENEVAL and the Examen Nacional para Aspirantes a Residencias Médicas (**ENARM**) include questions about Biostatistics?

The following is a conclusion of 130 practicing UK physicians that answered a research questionnaire:

"Grounding the teaching of statistics in the context of proper research studies and including examples of typical clinical work may better prepare medical students for their subsequent career".

How does Biostatistics make medical students better practitioners?

Virtually any medical research study uses Biostatistics from beginning to end. Medical students (including myself in my first years of med school) think Biostatistics is an unwarranted bore or academic burden. However, the stark reality is that practicing physicians recognize that **Biostatistics is important**. In fact, for those physicians who do not possess Biostatistics proficiency of their own, they rely on friends and colleagues. In medical practice, competence in Biostatistics is essential, is practical, is useful, and is used and applied—whether it is the physician's own knowledge or insights gathered from colleagues; or learned from other sources.

How do physicians know if the facts that are presented to them are true, or not so true? Unless they figure out the basics of Biostatistics, they may become prey to unscrupulous advertisers and industry promoters, or simply faulty research.

There may be bogus or biased findings that serve the vested interests of the research group, product company or sales force, but not necessarily fulfill the requirements of the physician or uphold the needs of the patient.

Now, what's the better way to get the medical students to **learn Biostatistics**?

Teaching Biostatistics should be targeted at the majority of the students, especially those students who will end up not conducting research. Students should be taught the unfamiliar language of statistics for the understanding of medical literature and communication with statistical consultants. They must learn to ask the right type of questions, rather than to apply recipes of mathematical methods. One thing that cannot be ignored is that medical students come to medical school to **become doctors, not statisticians**.

Data used in while trying to teach and learn Biostatistics should be relevant to the students' frame of reference, and current medical journal articles should be used to illustrate appropriate and inappropriate use of statistics. The way of examining Biostatistics should be adjusted to assess **insight** and not knowledge.

As the article "Teaching conceptual vs theoretical statistics to medical students" concludes: the approach to teaching statistics to preclinical medical students has been generally inaccurate. This course should focus on helping physicians understand the measurements and critically appraise the medical literature. It is way much better to teach conceptual statistics rather than in a theoretical way.

Bibliography and Suggested Reading

- Bahn AK. Teaching Biostatistics to Medical Students. *J Med Educ*. 1969;44 (7), 622-6.
- Dawson-Saunders B, Azen S, Greenberg RS, Reed AH. The Instruction of Biostatistics in Medical Schools. *The American Statistician* 1987. 41(4), 263-266.
- Feinstein AR. Clinical biostatistics; XXXIII. On teaching statistics to medical students. *Clin Pharmacol Ther*. 1975 Jul;18(1):121-6.
- Kerna NA. The relevance of biostatistics to the medical student. *Biom Biostat Int J*. 2018;7(2):87-88.
- Stander I. Teaching conceptual vs theoretical statistics to medical students. International Statistical Institute, 52nd session 1999; Helsinki, Finland.

Basic Principles of Biostatistics

Learning objectives for this chapter

- A. Recognize the fields of study of Statistics.
- B. Identify the potential areas of application of Statistics.
- C. Describe the roles biostatistics serves in public health and biomedical research.

Before we dive in some interesting stuff, please keep in mind one minor thing: **Statistics** are the tools used to **describe** and **analyze numbers in Medicine**.

Like all fields of learning, Statistics has its own vocabulary. Some concepts and phrases will be new while others, though appearing to be familiar, may have specialized meanings different from the definitions we already associate with these terms.

In a few words, we can describe Statistics as a field of study concerned with:

- » The **collection, organization, summarization, and analysis** of data.
- » The **drawing of inferences** about a universe of data when only a part of the data are observed.

When the data analyzed are derived from the biological sciences and Medicine, we use the term **Biostatistics** to distinguish this particular application of statistical tools and concepts.

Statistical methods include **procedures** for many things, such as:

1. Designing studies.
2. Collecting data.
3. Presenting and summarizing data.
4. Drawing inferences from sample data to a population.

These methods are useful in studies involving humans because the processes under investigation are often too intricate. Because of this complexity, many measurements on the study subjects are usually made to aid the discovery process; however, this intricacy and abundance of data usually mask the underlying processes.

It is in these situations that the systematic methods found in Statistics help **create order out of the chaos**.

Some **areas of application** of Statistics are:

- » Collection of vital statistics (e.g., mortality rates) used to **inform** about and to **monitor** the health status of the population.
- » Clinical trials to **determine** whether a new anti-hypertensive drug performs better than the standard treatment for mild to moderate essential hypertension.
- » Surveys to **estimate** the proportion of low-income women of child-bearing age with iron deficiency anemia.
- » Studies to investigate whether exposure to electromagnetic fields is a risk factor for leukemia.

Key Terms

Define the following terms.

Biostatistics

Statistics

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. **Answer:** What's the purpose of Statistics in Medicine?
2. **Answer:** As a field of study, what's Statistics concerned with?
3. **List the procedures included in statistical methods useful in studies involving humans.**
4. **List some areas of application of Biostatistics in your daily clinical practice.**

Bibliography and Suggested Reading

- Bowers D. Medical Statistics from Scratch. An Introduction for Health Professionals. 3rd Edition. West Sussex: John Wiley & Sons, Inc; 2014.
- Celis-De la Rosa AJ, Labrada-Martagón V. Bioestadística. 3^{ra} Edición. México: El Manual Moderno; 2014.
- Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3rd Edition. London: Elsevier Academic Press; 2012.
- Kerna NA. The relevance of biostatistics to the medical student. *Biom Biostat Int J.* 2018;7(2):87–88.
- Poldrack RA. Statistical Thinking for the 21st Century. Available at: <http://statstinking21.org>.
- Riffenburgh RH. Statistics in Medicine. 3rd Edition. Burlington: Elsevier Elsevier Academic Press; 2012.

Getting Down and Dirty with Data

Learning objectives for this chapter

- A. Explain how samples and populations, as well as a sample statistic and population parameter, differ.
- B. Distinguish between descriptive and inferential statistics.
- C. Distinguish between dependent and independent variables.
- D. Identify data relating to variables.
- E. Distinguish between qualitative and quantitative data.
- F. State the scales of measurements of variables and provide an example for each.
- G. Determine whether a value is discrete or continuous.
- H. Recognize that data and knowledge of Statistics allows you to investigate a wide variety of interesting phenomena.

What are Data?

The first important point about data is that data **are** - meaning that the word “data” is plural.

Data are composed of **variables**. A variable reflects a unique measurement or quantity.

Population

In Biostatistics, the population or universe is defined as the **set of values for which there is some interest**. The total of the universe or population is represented by the capital letter N.

Populations can be defined by determining a rule (or rules). These can be: characteristics of individuals, geographical boundaries, existing groups, time limits, etc. For example: residents of San Luis Potosí, students attending a school trip, IMSS right holders, cholera sufferers.

The elements of the universe can be people, places or things, whether they are **unique** or **grouped individuals**.



The **population** is defined as the set of values for which there is some interest.

The **population parameter** is the data obtained from the population.



The **sample** is a random subset of the population that is intended to represent the population.



The **sample statistic** is the data obtained from the sample.



For example: bedridden patients are elements that make up part of the universe defined as a hospital, but also the staff, students, furniture and different services provided in it can be elements of the same set.

When trying to study the entire population, the data obtained are referred to as **population parameter**. When researchers use data gathered by themselves, and **do not use them in any way**, that it, they merely describe the data, is called **descriptive statistics**.

Sample

Usually researchers in Clinical Investigation do not have the resources and time necessary to study every element in a population. That's why they study a **sample** of the population.

A sample is defined as a **subset of the population that is intended to represent the population**. Occasionally, the best way to get a sample that accurately represents the population is by selecting a **random** sample of the population, giving each individual in the population the **same chance** of being selected for the sample. The data collected from this sample are referred to as **sample statistic** and are used as an estimate of the population data (making an **inference**).

When researchers use a sample statistic to infer the value of a population parameter, it is called **inferential statistics**.

If the sample did not represent the population adequately, the sample statistic would NOT be similar to the population parameter. This would generate a **sampling error** (the difference between a sample statistic value and an actual population parameter value).

Some differences to recognize between both types of Statistics are shown in **Table 8.1**.

Independent and Dependent Variables

Experiments are designed to test if one or more variables **cause modifications** to another variable.

For example, if a researcher considers that a new treatment reduces depressive symptoms, he could design an experiment to test this prediction. He might give a sample of people with depression the new treatment and withhold the treatment from another sample of people with depression. Later, if those who received the new treatment had lower levels of depression, he would have evidence that the new treatment reduces depression.

» In this experiment, the type of treatment each person received (i.e., new treatment vs. no treatment) is the **independent variable**.

Table 8.1. Key Differences Between Descriptive Statistics and Inferential Statistics

Descriptive Statistics	Inferential Statistics
Concerned with describing the population	Makes inferences from the sample and generalize them to the population
Organize, analyze, and present the data in a meaningful manner	Compares, tests and predicts future outcomes
Final results are shown in forms of charts, tables and graphics	Final results are shown as probability scores
Describes the data already known	Tries to make conclusions about the population that is beyond the data available
Tools: measures of central tendency (mean/median/mode), spread of data (range, standard deviation)	Tools: hypothesis test, analysis of variance, etc.
Usually there is less error involved	Usually there is more error involved

» In this experiment, the number of depressive symptoms observed in each person is the **dependent variable**.

More generally:

» The **independent variable** is a variable with two or more levels that are expected to have different impacts on another variable.

» The **dependent variable**, on the other hand, is the outcome variable that is used to compare the effects of the different independent variable levels.

In true experiments, the **manipulated variable** (by the investigators) is always referred to as the **independent variable**.

Independent variable causes a **change** in the **Dependent Variable**. It isn't possible that the **Dependent Variable** could cause a change in the **Independent Variable**.

Operational Definition of Variables

All the variables used in any statistical work must be **clearly defined**, in order to avoid confusion, facilitate the search and analysis of the data, and guarantee the comparability of the results with those obtained in other studies. This is particularly important when variables can be defined in different ways.

Types of Variables

The variables can be classified into **qualitative** and **quantitative** (Figure 8.1).

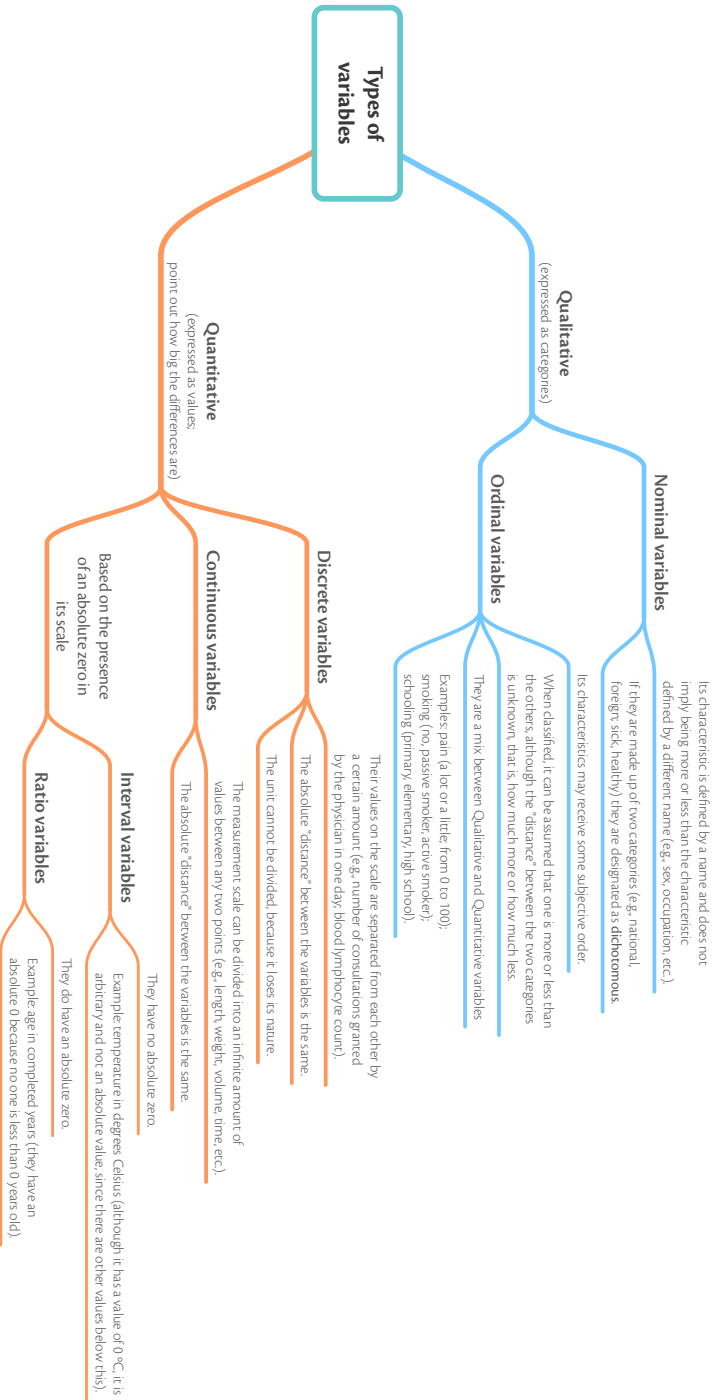


Figure 8.1. Diagram of the types of variables.

Simplifying:

- » When their characteristics are expressed as **categories**, they are said to be **qualitative variables**.
- » When they are expressed as **values**, they are identified as **quantitative variables**.

Qualitative data

When variables are qualitative, it means that they describe a **quality** rather than a numeric quantity.

Take the next question as an example. “What is your favorite food?” All the answers (data) we could collect are not intrinsically numerical. We could assign numbers to each one (1 = pizza, 2 = chocolate, etc), but we would just use those numbers as labels rather than as real numbers. It wouldn’t make sense to add the numbers together.

Quantitative data

Quantitative data are more commonly used in Biostatistics. These data are **numerical**.

For example, let’s take a look at the data generated at an undergraduate Statistics course at Stanford, where the professor asked his students “Why are you taking this class?”. The results are summarized in **Table 8.2**.

Although the students’ answers were qualitative, a quantitative summary of data was generated by counting how many students gave each response.

Types of Numbers

There are several different types of numbers that are used in Biostatistics:

- » **Binary numbers:** are the simplest numbers – that is, zero or one. They are used to represent whether something is true or false, or present or absent.

Table 8.2. Data From the Example on Quantitative Data

“Why are you taking this class?”	Number of Students
It fulfills a degree plan requirement	105
It fulfills a General Education Breadth Requirement	32
It is not required but I am interested in the topic	11
Other	4

» **Integers:** are whole numbers with no fractional or decimal part. They are used when counting things and in psychological measurements (“Disagree Strongly” - “Agree Strongly”).

» **Real numbers:** are the most common numbers used in Biostatistics. They have a fractional/decimal part (e.g., when you measure someone’s weight, it can be measured to an arbitrary level of precision, from whole pounds down to micrograms).

Scales of Measurements

Remember that all research is based on **measurement**. The way a variable is measured has a direct impact on the types of statistical procedures that can be used to **analyze** that variable.

Generally speaking, researchers want to devise measurement procedures that are as precise as possible because more precise measurements enable more sophisticated statistical procedures. That’s why there are **four different scales of measurement** recognized that vary in their degree of measurement precision (**Figure 8.1**):

1. **Nominal:** Categorize things into groups that are qualitatively different from other groups.
 - Therefore, they yield **qualitative data**.
2. **Ordinal:** Categorize things into different groups, but on ordinal scale, that is, differ the amount of something they possess.
 - Therefore, they yield **qualitative data**.
3. **Interval:** Indicate exactly how much of something people have.
 - Therefore, they yield **quantitative data**.
4. **Ratio:** Involve quantifying how much of something people have, but a score of zero indicates that the person has none of the thing being measured.
 - Therefore, they yield **quantitative data**.

Each of these scales of measurement is increasingly more precise than its predecessor, and therefore allows the performance of more sophisticated statistical analyses.

Discrete vs. Continuous Measurements

A **discrete** measurement **takes one of a set of particular values**.

These could be qualitative values (e.g., different breeds of dogs) or numerical values (e.g., how many friends one has on Facebook).

» **Important:** there is no middle ground between the measurements; it doesn't make sense to say that one has 33.7 friends on Facebook.

A **continuous** measurement is **defined in terms of a real number**. It could fall anywhere in a particular range of values, though usually our measurement tools will limit the precision of its measure (e.g., a floor scale might measure weight to the nearest pound, even though weight could in theory be measured with much more precision).

How Can I Tell the Type of Variable I Am Dealing With?

The easiest way to tell whether data are **quantitative** is to analyze if it has **units attached**, such as g, mm, °C, $\mu\text{g}/\text{cm}^3$, number of pressure sores and number of deaths. If not, it may be **ordinal** or **nominal** –the former if we can put the values in any meaningful order.

Figure 8.2 presents an algorithm that can assist to variable-type recognition.

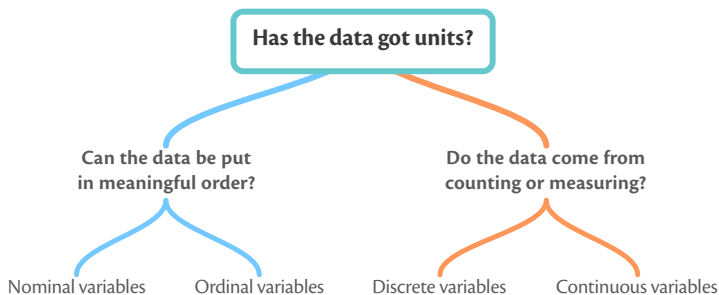


Figure 8.2. Algorithm to identify data type.

Key Terms

Define the following terms.

Binary numbers	Integers	Quantitative data
Continuous measurement	Interval	Ratio
Data	Nominal	Real numbers
Dependent variable	Operational definition of a variable	Sample
Descriptive statistics	Ordinal	Sample statistic
Discrete measurement	Population	Sampling error
Independent variable	Population parameter	Scale of measurement
Inferential statistics	Qualitative data	Variable

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Complete the following sentences with the word or short phrase that has been left out.

- In statistics the term we use for a set of figures is _____.
- A sample is part of a _____.
- A kind of Statistics used for estimating population characteristics is known as _____.

2. **Table AL8.1** shows a portion of a data set that was created to collect data for a study on risk factors for myocardial infarction (MI), the primary outcome of the study. Eligible individuals who had not previously experienced any cardiovascular disease (including MI) were recruited from January to December 2000 and were males and females aged 30–65 years. Information was collected on a total of 1113 individuals; only data for the first 30 individuals are shown in **Table AL8.1**.

- For each variable in the data set, identify the type of variable.

Table AL8.1. Data From the Study

Variable	Variable Description
Entry date	Date of recruitment in study
Sex	M = Male, F = Female
Age	Age in years
Ethnic group	1 = White; 2 = Black African; 3 = Other black; 4 = Other; 9 = Not known
Smoking	0 = Never smoker; 1 = Current smoker; 2 = Ex-smoker; 9 = Not known
BMI	Body mass index (kg/m ²)
TC	Total cholesterol (mmol/L)
HDL-C	High-density lipoprotein cholesterol (mmol/L)
TG	Triglyceride (mmol/L)
LLD	Any receipt of lipid-lowering drugs (Y = yes; N = no)
Date of first MI	Date of first MI during study

- List the error checks that you would perform on each variable – are there any entries that you would wish to investigate further?

3. In a study on the effects of coffee drinking on cardiovascular health, it was found that drinking coffee was associated with a greater risk for heart disease. Upon further analysis, researchers discovered that a large majority of coffee drinkers were smokers and that smoking status was the true cause of the apparent association. Based on this case:

- Determine the dependent variable.
- Determine the independent variable.

4. A researcher studies how different drug doses affect the progression of a disease and compares the intensity and frequency of symptoms when different doses are given. Based on this case:

- Determine the dependent variable.
- Determine the independent variable.

5. Many continuous variables are dichotomized to make them easier to understand e.g. obesity (body mass index >30 kg/m²) and anemia (hemoglobin level <10 g/dl). Based on this scenario, answer the following statements:

- What information is lost in this process?
- If you were told that a patient was anemic, what further information would you want to know before treating the patient?
- How does a label, such as “anemia”, would help?

6. A physician in an emergency room (ER) is collecting data. What sort of data are the following:

- Time in minutes waiting in ER.
- Triage outcome (no injury, minor injury, major injury).
- Number of cases of road accident victims in the ER.

7. Multiple choice questions.

1. If a sample represents a population well, it will:

- Respond in a way that is similar to how the entire population would respond.
- Generate a large amount of sampling error.

2. Which one of the following statements is true?

- A qualitative variable comprises two categories which may be ordinal or numerical.
- An ordinal variable comprises categories which cannot be ordered.
- The age groups ‘young’, ‘middle aged’ and ‘old’ relate to a nominal categorical variable.
- Blood group is classified as a nominal categorical variable.
- It may be difficult to distinguish a continuous numerical variable from an ordinal variable when the ordinal variable has many categories.

3. The value obtained from a population is called:

- Statistic.
- Parameter.

4. Parameters are:

- Always exactly equal to sample statistics.
- Often estimated or inferred from sample statistics.

5. When a statistic and parameter differ,

- There is sampling error.
- It is called an inferential statistic.

6. Researchers are using descriptive statistics if they are using their results to:

- Estimate a population parameter.
- Describe the data they actually collected.

7. Researchers are using inferential statistics if

they are using their results to:

- a) Estimate a population parameter.
- b) Describe the data they actually collected.

8. The IV (independent variable) in a study is the:

- a) Variable expected to change the outcome variable.
- b) Outcome variable.

9. The DV (dependent variable) in a study is the

- a) Variable expected to change the outcome variable.
- b) Outcome variable.

10. All studies allow you to determine if the IV causes changes in the DV.

- a) True.
- b) False.

11. The way a variable is measured:

- a) Determines the kinds of statistical procedures that can be used on that variable.
- b) Has very little impact on how researchers conduct their statistical analyses.”

12. Researchers typically treat summed questionnaire/survey scores as which scale of measurement?

- a) Nominal scale of measurement.
- b) Ordinal scale of measurement.
- c) Interval scale of measurement.

13. The scale of measurement that quantifies the thing being measured (i.e., indicates how much of it there is):

- a) The nominal.
- b) The ordinal.
- c) Both the interval and ratio.

14. The scale of measurement that categorizes objects into different kinds of things is:

- a) The nominal.
- b) The ordinal.
- c) Both the interval and ratio.

15. The scale of measurement that indicates that some objects have more of something than other objects but not how much more is:

- a) The nominal.
- b) The ordinal.
- c) Both the interval and ratio.

16. If a variable can be measured in fractions of units, it is a: _____ variable.

- a) Discrete.
- b) Continuous.

17. You are conducting an experiment to see if exposure to more sunlight increases happiness levels for workers who typically spend the entire day in windowless offices. Choose the dependent variable.

- a) Sunlight.
- b) Time of day.
- c) Windowless offices.
- d) Happiness level.

18. A researcher suspects that a cholera outbreak is happening because of tainted wells in the city. Most of the cases are clustered around public wells that draw their water from the underground aquifer. Choose the independent variable.

- a) The underground aquifer.
- b) Wells.
- c) Cholera.
- d) The City.

19. Studies have shown that condom use is effective in controlling the spread of HIV. However, studies also show that a combination of two HIV medications (tenofovir and emtricitabine) can also control the spread of the disease. Choose the dependent variable.

- a) Tenofovir.
- b) Emtricitabine.
- c) Both a) and b).
- d) HIV.

20. In an experiment, the variable that is being measured is referred to as the:

- a) Independent variable.
- b) Confounding variable.
- c) Dependent variable.
- d) Measurement variable.
- e) Dependant variable.

Bibliography and Suggested Reading

- Bowers D. Medical Statistics from Scratch. An Introduction for Health Professionals. 3rd Edition. West Sussex: John Wiley & Sons, Inc; 2014.
- Celis-De la Rosa AJ, Labrada-Martagón V. Bioestadística. 3^{ra} Edición. México: El Manual Moderno; 2014.
- Esparza-Villalpando V. Basic Biostatistics with R. MCIC. San Luis Potosí; 2019.
- Kaliyadan F, Kulkarni V. Types of Variables, Descriptive Statistics, and Sample Size. Indian Dermatol Online J. 2019 Jan-Feb; 10(1):82-86.
- Pezzullo JC. Biostatistics for Dummies. Hoboken, NJ: John Wiley & Sons, Inc; 2013.
- Poldrack RA. Statistical Thinking for the 21st Century. Available at: <http://statstinking21.org>.

Descriptive Statistics

Learning objectives for this chapter

- A.** Identify appropriate numerical and graphical summaries for each variable type.
- B.** Understand data presented in a graph and organize data into frequency distribution graphs, including bar charts, pie charts, histograms, and Box-and-whiskers plot.
- C.** Understand the concept of a frequency distribution as an organized display showing where all of the individual scores are located on the scale of measurement.
- D.** Understand the purpose of measuring central tendency and identify the circumstances in which its use is appropriate.
- E.** Define and interpret summary statistics for a quantitative variables, including mean, median, standard deviation, range, and IQR.
- F.** Understand the relation between the measures of central tendency in symmetrical and skewed distributions.
- G.** Learn the characteristics and properties of a normal curve.
- H.** Understand standard error and variability.
- I.** Understand the concept of the Central Limit Theorem and its application to increase the accuracy of measurements.
- J.** Recognize that a confidence interval will capture the true parameter for the specified percentage of all random samples and interpret a confidence interval in context.

Descriptive Statistics are required to **summarize large data sets**, so they can be clearly illustrated.

The properties of a parameter are specified by their so-called **scale of measure**. As we studied in the last chapter, generally two types of parameters are distinguished:

- » A variable may have a metric level (**quantitative data**) if it can be counted, measured or weighted in a physical unit or at least can be recorded in whole numbers.
- » Likewise, a variable may have a category classification (**qualitative data**) if they cannot be measured, but classified.

Some types of data are best described with a table, some with a chart and some perhaps with both, whereas with other types of data, a numeric summary might be more appropriate. In this chapter we will learn how and when to use these resources.

Summarizing and Graphing Categorical Data

A categorical variable is summarized in a fairly straightforward way. You just tally the number of subjects in each category and express this number as a count—and perhaps also as a percentage of the total number of subjects in all categories combined.

To better illustrate this concept, as well as the following ones, we will take a sample of 422 subjects. This sample will be summarized by race, as is shown in **Table 9.1**.

The joint distribution of subjects between two categorical variables (such as Race by Gender), can be summarized by a **cross-tabulation** (“cross-tab”), as is shown in **Table 9.2** with the same sample of 422 subjects.

Categorical data are usually summarized graphically as **frequency bar charts** or as **pie charts**.

Frequency Bar Charts

Displaying the spread of subjects across the different categories of a variable is most easily done with a **bar chart** (**Figure 9.1-A**).

To create a bar chart manually from a tally of subjects in each category, you draw a graph containing **one vertical bar for each category**, making the height proportional to the number of subjects in that category.

Pie Charts

This resource shows the relative number of subjects in each category by the angle of a circular wedge (like a piece of the pie) (**Figure 9.1-B**).

To create a pie chart manually, you multiply the percent of subjects in each category by 360 (the number of degrees of arc in a full circle), and then divide by 100.

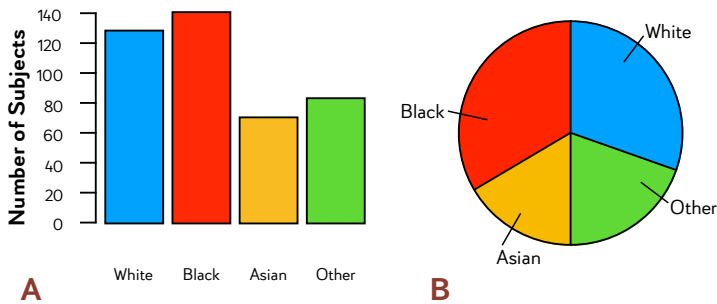
Special considerations:
Comparing the relative magnitude of the different sections of a pie chart is more difficult than comparing bar heights.

Table 9.1. Example of Subjects Categorized by Race

Race	Count	Percent of Total
White	128	30.3%
Black	141	33.4%
Asian	70	16.6%
Other	83	19.7%
Total	422	100%

Table 9.2. Example of a Cross-Tab of Subjects by Two Categorical Variables

	White	Black	Asian	Other	Total
Male	60	60	34	42	196
Female	68	81	36	41	226
Total	128	141	70	83	422

**Figure 9.1.** Graphic examples for frequency bar charts and pie charts.

Summarizing Numerical Data

Numerical variables aren't as simple to summarize as categorical variables.

The summary statistics for a numerical variable should convey, in a concise and meaningful way, how the individual values of that variable are **distributed across your sample of subjects**, and should give you some idea of the **shape of the true distribution** of that variable in the population from which you draw your sample.

That true distribution can have almost any shape, including the typical shapes shown in **Figure 9.2**: normal, skewed, pointy-topped, and bimodal (two-peaked).

Frequency distributions have four important characteristics:

- » **Center:** Where do the numbers tend to center
- » **Dispersion:** How much do the numbers spread out?
- » **Symmetry:** Is the distribution shaped the same on the left and right sides, or does it have a wider tail on one side than the other?
- » **Shape:** Is the top of the distribution nicely rounded, or pointier or flatter?

These characteristics are measured using **numbers**, which will be detailed below.

Center

It's perhaps the most important single thing that needs to be known about a set of data: at what value the data tend to center around. This characteristic is called **central tendency**, and three measures of central tendency are described: **mean**, **median**, and **mode**.

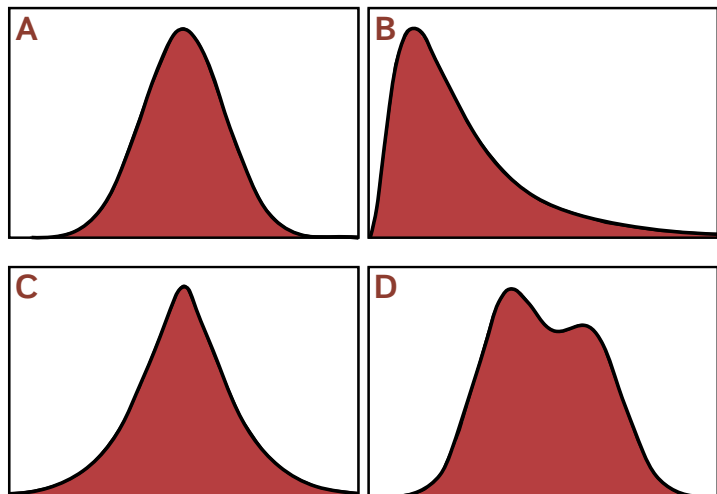


Figure 9.2. Graphic examples of four different shapes of distributions: normal (a), skewed (b), pointy-topped (c), and bimodal (two-peaked) (d).

Arithmetic mean

Commonly called the mean or the **average**, is the most familiar and most often used measure of central tendency.

It is only a number that summarizes a series of values from which it is calculated. You can obtain it by adding all the values of a certain population and dividing the total by the number of values that were added.

It is represented symbolically by the Greek letter μ (mu) when it is obtained from **population** data, and by \bar{x} when it is estimated from a random **sample**.

The mean is the mathematical result that **synthesizes the data in a single figure**, and we must not forget that it only **describes the group as such** and not each of its elements.

The following are **properties of the mean**:

- » **Uniqueness**: for a given set of data, there is only one arithmetic mean.
- » **Simplicity**: the arithmetic mean is easy to understand and calculate.
- » In the data series, **all values are used** for its calculation. Therefore, extreme values can skew the result.

The mean is used to summarize **quantitative data** when the study group is **too large** or when the series of observations has **no extreme values**.

Median

Is defined as that value that is **in the middle of a population** whose values are **ordered according to their magnitude**.

If the number of observations is odd, the median will be the value that lies in between. When the number of observations is even, the average of the two observations in between is taken.

The median can be obtained as follows:

1. The values of the variable are ordered from least to greatest and are numbered progressively.
2. The middle value is determined by $0.5(N + 1)$, regardless of whether N is even or odd.
3. If the previous equation provides an integer, the median value corresponds to the one in that position. Otherwise, the fraction that follows the integer is multiplied by the difference between the two ordered values of the variable and the result is added to the smaller value.

The following are **properties of the median**:

- » Unique.
- » Simple.
- » Extreme values do not affect it (as in the mean).
- » Divides the data into two equal parts, each with 50% of the observations.

The following are **disadvantages** of the median in relation to the mean:

- » **Disdain information**, because it only considers the values of 1 or 2 observations.
- » When two or more groups come together in one, you cannot calculate a median from the median of each group.

The **median** is used to summarize **quantitative data** when the group under study is **small** and **does not have a symmetric distribution**.

Mode

Is the **most repeated value in a group of data**.

A group of data can have more than one mode. This measure can be used for both qualitative and quantitative variables. However, its use is **limited** because of the little information it provides.

Dispersion

The second most important thing to understand about a set of numbers is **how tightly or loosely they tend to cluster around a central value**; that is, how narrowly or widely they're dispersed.

There are several common ways to measure this dispersion.

Standard Deviation (SD or sd)

Tells you **how much the individual numbers tend to differ from the mean** (in either direction).

When talking about population distributions, the SD describes the width of the distribution curve. **Figure 9.3** shows three normal distributions. They all have a mean of zero, but they have different standard deviations and therefore, different widths. Each distribution curve has a total area of exactly 1.0, so the peak height is smaller when the SD is larger.

Standard deviations are **very sensitive to extreme values** (outliers) in the data.

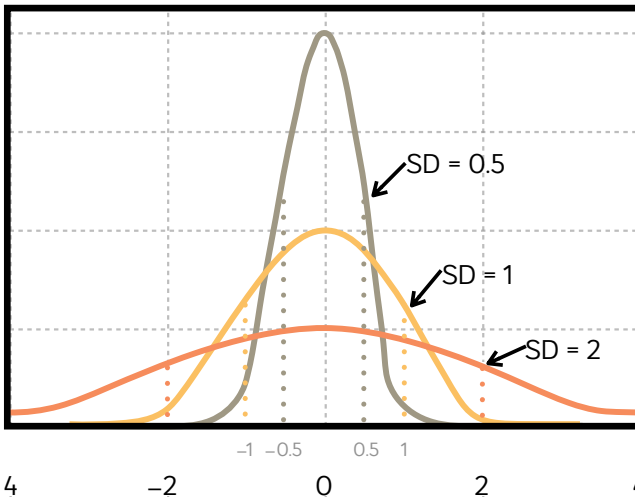


Figure 9.3. Three distributions with the same mean but different standard deviations.

Several other useful measures of dispersion are related to the SD:

- » **Variance:** Is just the square of the SD.
- » **Coefficient of variation (CV):** Is the SD divided by the mean.

Range

Is the **difference** between the smallest value (the minimum value) and the largest value (the maximum value):

$$\text{Range} = \text{maximum value} - \text{minimum value.}$$

As such, it is **extremely sensitive to outliers**.

Centiles

The basic idea of the median (that half of your numbers are below the median) can be extended to other fractions besides $1/2$.

A **centile** is a value to which a certain **percentage** of the values are below. For example, $1/4$ of the values are less than the 25th centile (and $3/4$ of the values are greater).

The median is just the **50th centile**. Some centiles have common nicknames:

- » The 25th, 50th, and 75th centiles are called the first, second, and third **quartiles**, respectively.
- » The 20th, 40th, 60th, and 80th centiles are called **quintiles**.

» The 10th, 20th, 30th, and so on, up to the 90th centile, are called **deciles**.

The **inter-quartile range** (or IQR) is the **difference between the 25th and 75th centiles** (the first and third quartiles).

When summarizing data from strangely shaped distributions (aka, **not normal**), the **median** and **IQR** are used instead of the mean and SD.

Symmetry and Shape

Skewness

Refers to whether the distribution has **left-right symmetry** or whether it has a **longer tail on one side or the other** (Figure 9.4).

Many different skewness coefficients have been proposed over the years; the most common one, often represented by the Greek letter γ (lowercase gamma), is calculated by averaging the cubes (third powers) of the deviations of each point from the mean and scaling by the standard deviation. Its value can be positive, negative, or zero.

- » A **negative** skewness coefficient (γ) indicates **left-skewed data** (long left tail).
- » A **zero** γ indicates **unskewed data**.
- » A **positive** γ indicates **right-skewed data** (long right tail).

Kurtosis

Is a way of quantifying the **differences in the shape of the distributions**.

There are three basic distributions (Figure 9.5):

- » A pointy top, and fat tails (**leptokurtic**).
- » Normal appearance.
- » Broad shoulders, small tails, and not much of a pointy top (**platykurtic**).

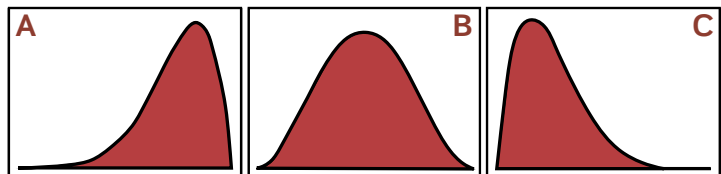


Figure 9.4. Distributions can be left-skewed (a), symmetric (b), or right-skewed (c).

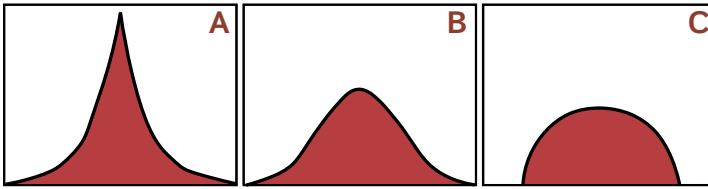


Figure 9.5. Three distributions: leptokurtic (a), normal (b), and platykurtic (c).

Gathering Together Summaries and Descriptive Tables

Descriptive tables contain the **most useful summary statistics** of the results and are arranged in a concise way.

Some of the biostatistical reports include N, mean, SD, median, minimum and maximum, and are arranged like this:

- » Mean \pm SD (N).
- » Median (minimum - maximum).

The real utility of this summary is that it allows you to identify **changes over time and between groups**. **Table 9.3** shows an example of this summary resource.

Putting Numerical Data into Graphics

Displaying information graphically is a central part of interpreting and communicating the results of a scientific research. Besides, by graphing our data we can easily **spot unnoticed subtle features** in a table of numbers.

Table 9.3. Example of a Descriptive Table

Systolic Blood Pressure Treatment Results		
	Drug	Placebo
Before Treatment	138.7 \pm 10.3 (40)	141.0 \pm 10.8 (40)
	139.5 (117–161)	143.5 (111–160)
After Treatment	121.1 \pm 13.9 (40)	141.0 \pm 15.4 (40)
	121.5 (85–154)	142.5 (100–166)
Change	-17.6 \pm 8.0 (40)	-0.1 \pm 9.9 (40)
	-17.5 (-34–4)	1.5 (-25–18)

Histograms

Histograms are **bar charts** that show what fraction of the subjects have values falling within specified intervals.

The main purpose of a histogram is to show you **how the values of a numerical value are distributed**. This distribution is an approximation of the true population frequency distribution for that variable (**Figure 9.6**).

Because a sample is only an imperfect representation of the population, determining the precise shape of a distribution can be difficult unless the sample size is very large. Nevertheless, a histogram usually helps to **spot skewed data**.

The kind of shape that occurs very often in Medicine is typical of a **log-normal distribution**. It's called "log-normal" because if you take the logarithm of each data value (it doesn't matter what kind of logarithm you take), the resulting logs will have a normal distribution (more on the normal distribution in a few lines ahead).

Log-normality isn't the only kind of non-normality that can arise in real-world data. Depending on the underlying process that gives rise to the data, the numbers can be distributed in other ways.

There are many ways to transform the data to make them look "approximately normal". However, when failing to accomplish it, we must analyze the data using **nonparametric methods**, with which we don't assume that the data are normally distributed.

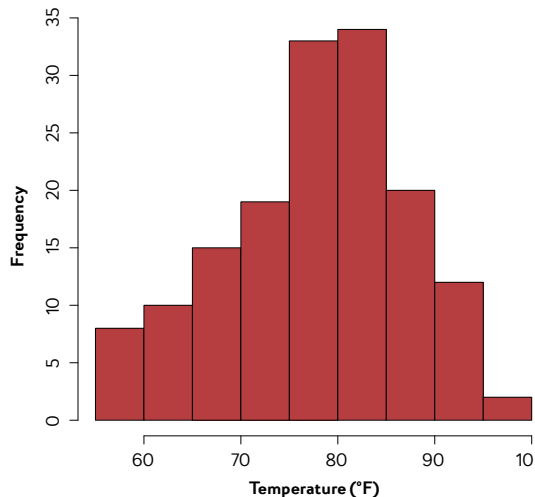


Figure 9.6. Histogram showing the frequency of the temperature of the air quality measurements in LaGuardia Airport, New York, from May to September 1973.

Bars, Boxes, and Whiskers

If you want to show how a variable **varies from one group of subjects to another**, two types of graphs are commonly used for this purpose: bar charts and box-and-whiskers plots.

Bar Charts

This graph allows us to display and **compare the means of several groups of data** (Figure 9.7).

The bar height equals the mean (or median) value of the variable.

The **error bars** are lines that represent the spread of values for each variable (SD above and below the top value of the bar).

Cons: it doesn't give a very good picture of the distribution of the variable within each group nor gives information about the skewness.

Box-and-Whiskers Plots

Also called **Box plot**. This graph represents a lot of information about the **distribution of numbers in one or more groups** of subjects (Figure 9.8).

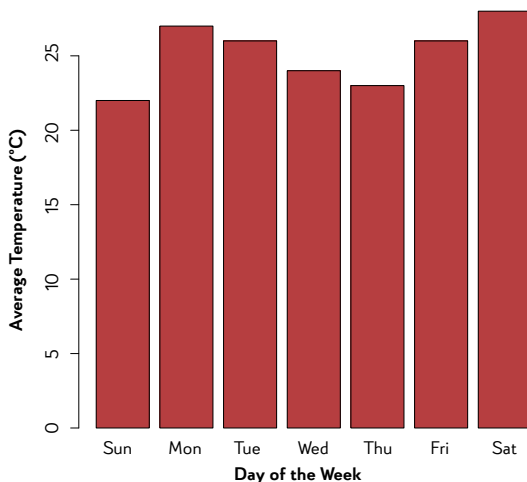


Figure 9.7. Bar chart showing the average by day of the temperature of the air quality measurements in LaGuardia Airport, New York, from May to September 1973.

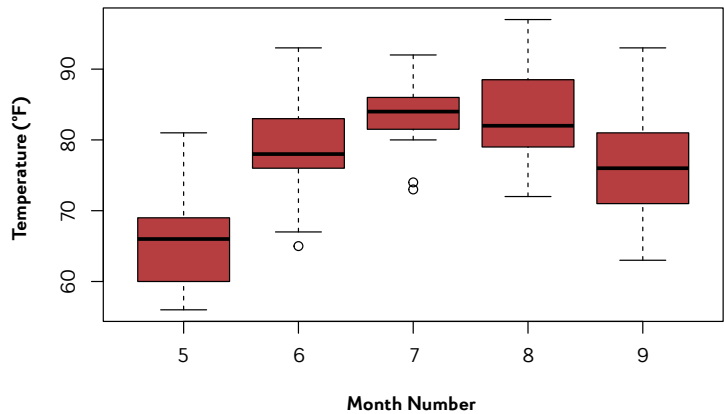


Figure 9.8. Box-and-whiskers plot showing the temperature of the air quality measurements in LaGuardia Airport, New York, from May to September 1973.

A Box plot has the following components:

- » A **box** spanning the **interquartile range (IQR)**.
 - Extends from the first quartile (25th centile) to the third quartile (75th centile) of the data, encompassing the middle 50% of the data.
- » A thick **horizontal line**, drawn at the **median** (50th centile), usually located at or near the middle of the box.
- » **Dashed lines** (whiskers) extending out to the farthest data point that's not more than 1.5 times the IQR away from the box.
- » **Individual points** lying outside the whiskers, considered **outliers**.

Useful tip:

A median that's not located near the middle of the box indicates a skewed distribution.

The Normal Distribution

The normal distribution is simply a distribution with a certain shape. It is normal because many things have this same shape. Is a specific frequency distribution pattern that is common in biological data, for which many statistical tests have been designed (e.g. t-test, analysis of variance).

Normal distributions are **bell-shaped and symmetric**. The mean, median, and mode are **equal**, and the variability is described by the **standard deviation**. The characteristics of a normal distribution are illustrated in **Figure 9.9**.

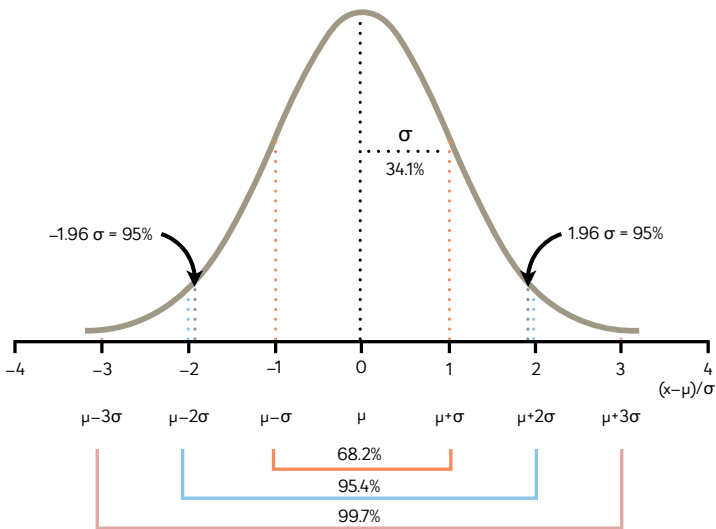


Figure 9.9. Characteristics of the Normal Distribution.

The **empirical rule** for the Normal Distribution (also known as the 68-95-99.7 rule) states:

- » About **68%** of the values in a normal distribution are **within one standard deviation of the mean**.
- » About **95%** of the values in a normal distribution are **within two standard deviations of the mean**.
- » About **99.7%** of the values in a normal distribution are **within three standard deviations of the mean**.

It is easy to re-create any normal distribution if you know two parameters: the **mean** and the **standard deviation**. The mean is the center of the bell-shaped picture, and the standard deviation is the distance from the mean to the inflection point (the place where the concavity of the curve changes on the graph).

The Standard Error

If you take a random sample from a population and calculate the sample mean, this information provides with an estimate of the population mean. However, if we analyze a different sample, it may give a different estimate of the population mean.

So if we take (say) 100 samples all of the same size, $n = 4$, we would get a spread of sample means which we can display visually in a histogram. The **variability** of these sample means gives us

an indication of the **uncertainty** attached to the estimate of the population mean when taking only a single sample (very uncertain when the sample size is small to much less uncertainty when the sample size is large).

The standard error or **variability of the sampling distribution of the mean** is measured by the standard deviation of the estimates. If we know the population standard deviation, s , then the standard error of the mean is given by $\sigma_{\text{sqd}/n}$.

Fundamental Knowledge

- » The mean of all the sample means will be the same as the population mean.
- » The standard deviation of all the sample means is known as the **standard error (SE)** of the mean or SEM.
- » Given a large enough sample size, the distribution of sample means, will be **roughly Normal** regardless of the distribution of the variable.

The standard error is a measure of the **precision of a sample estimate**. It provides a measure of how far from the true value in the population the sample estimate is likely to be.

All standard errors have the following interpretation:

- » A **large** standard error indicates that the estimate is **imprecise**.
- » A **small** standard error indicates that the estimate is **precise**.
- » The standard error is **reduced**, that is, we obtain a more precise estimate, **if the size of the sample is increased**.

The Central Limit Theorem

The Central Limit Theorem states that **the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger**—no matter what the shape of the population distribution.

That statement means that as you take more samples, especially large ones, the graph of the sample means will look more like a normal distribution.

An essential component of the Central Limit Theorem is that **the average of the sample means will be close to the population mean**. In other words, add up the means from all of your samples, find the average and it will be close to your actual population mean. The same concept applies to the standard deviations. It's a pretty useful phenomenon that can help accurately **predict characteristics of a population**.

Easily explained:

Basically, the Central Limit Theorem is saying that the more times you roll a die, the more likely the shape of the distribution of the means tends to look like a normal distribution graph.

Confidence Intervals

Confidence Intervals (CI) indicate a **range of values that's likely to encompass the truth**, that is, a range of values that **surround the sample statistic** in order to correspond with the value of the population parameter.

CI are written as a pair of numbers separated by a dash, like this: 114-126. The two numbers make up the lower and upper boundaries (**confidence limits**).

The probability that the confidence interval encompasses the true value is called the **confidence level of the confidence intervals**. The most commonly value is **95%**, that's why whenever a researcher reports a confidence interval, he states the confidence level like this: 95%CI = 114-126.

As a **general rule**: higher confidence levels correspond to wider confidence intervals, and lower confidence level intervals are narrower.

Properly calculated, 95% confidence intervals contain **the true value 95% of the time** and fail to contain the true value the other 5% of the time. Usually, this confidence limits are calculated to be balanced so that the 5% failures are split evenly. If the data are normally distributed, with a mean of 0 and a standard deviation of 1, the confidence intervals would be **-1.96** and **+1.96**, because 95% of all z-scores fall within these values (**Figure 9.9**).

To calculate the confidence intervals with large samples (higher than 30), you can use the following equations:

- » **Lower boundary:** $\bar{x} - (1.96 * SE)$.
- » **Upper boundary:** $\bar{x} + (1.96 * SE)$.

You can also use the CI as an alternative to assess **significance**. In order to do so, you first define a number that measures the amount of effect you're testing for. This effect size can be the difference between two means or two proportions, the ratio of two means, an odds ratio, a relative risk ratio, or a hazard ratio, among others (more on this in the following chapters). The complete absence of any effect (no-effect value) corresponds to a difference of 0, or a ratio of 1. Based on this:

- » If the 95% CI around the observed effect size includes the no-effect value (0 for differences, 1 for ratios), then the effect is **not statistically significant** (that is, a significance test for that effect will produce $p > 0.05$).
- » If the 95% CI around the observed effect size does not include the no-effect value, then the effect is **statistically significant** (that is, a significance test for that effect will produce $p \leq 0.05$).

Easily explained:
Confidence Intervals are an indicator of the precision of a numerical quantity focused on the population.

The same kind of correspondence is true for other confidence levels and significance levels:

- » 90% confidence levels correspond to the $p = 0.10$ significance level.
- » 99% confidence levels correspond to the $p = 0.01$ significance level, and so on (more on the p -value in the following Chapters).

Key Terms

Define the following terms.

Bar chart	Frequency distribution	Qualitative data
Box-and-whiskers plot	Histogram	Quantitative data
Center	Inter-quartile range	Range
Centiles	Kurtosis	Scale of measure
Central limit theorem	Mean	Shape
Central tendency	Median	Significance
Confidence interval	Mode	Skewness
Confidence limits	Nonparametric method	Standard deviation
Cross-tabulation	Normal distribution	Standard error
Descriptive statistics	Parametric method	Statistically significant
Descriptive table	Pie chart	Symmetry
Dispersion	Properties of the mean	Uncertainty
Frequency bar chart	Properties of the median	Variability

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Based on the data shown in Table AL8.1:

- Describe the summary measures (if any) that you would use to summarize the variable.
- What graphical methods would you use to display each of the variables?

2. Make a sketch of the following, indicating the approximate locations for the mean, median and mode:

- A normal distribution.
- A skewed distribution.
- A rectangular distribution.

3. The following numbers represent the time in minutes that twelve medical students took to get to school on a particular day. 18, 34, 68, 22, 10, 92, 46, 52, 38, 29, 45, 37

- Calculate the quartiles and find the interquartile range.

4. Histograms of grades (out of 100) on three different exams for a group of 150 students are given in Figure AL9.1. The pass grade for each exam is 6.0.

- For each exam, was the percentage who passed about 50%, well over 50% or well under 50%?

5. Based on the box-plot shown in Figure 9.8:

- Comment on the differences in the shape, spread, and location of the box-plots.
- Calculate the median, the first quartile and the third quartile for the data of June.
- Calculate the interquartile range for the data of August.
- Identify and determine the outliers present in the graph.

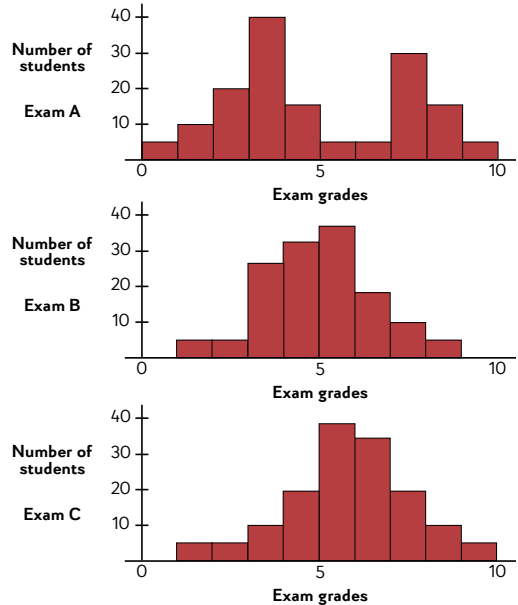


Figure AL9.1. Histograms of grades on 3 exams.

6. Individuals with cystic fibrosis (CF) are subject to recurrent respiratory infections (exacerbations) that often require intravenous antibiotic treatment and may result in permanent loss of lung function. In 2010, Collaco et al. performed a retrospective study on 1535 subjects in the United States to investigate the effects of delivering therapy in hospital and at home. Antibiotic treatment was given following an exacerbation. When raw forced expiratory volume (FEV1) measurements were converted to Knudson percentiles, the authors found that long-term decline in FEV1 after exacerbation was observed regardless of whether antibiotics were administered in the hospital (mean = -3.3 and SD = 8.4 percentage points for $n = 602$ courses of therapy) or at home

(mean = -3.5 and SD = 7.6 percentage points for $n = 232$ courses of therapy). No significant difference in intervals between courses of antibiotics was observed between hospital (median 119 days, interquartile range 166 (221 – 55) days) and home (median 98 days, interquartile range 155 (204 – 49) days) ($P = 0.29$).

- Determine the 95% confidence interval for the mean decline in FEV1 (expressed in percentage points) for both the hospital and home administration of antibiotics.
- Interpret the 95% confidence interval for the mean decline in FEV1 (expressed in percentage points) for the hospital administration of antibiotics.
- Why do you think the authors provided the median interval between courses of antibiotics rather than the mean interval?

7. In the campaign against smallpox, a doctor inquired into the number of times 150 people aged 16 and over in an Ethiopian village had been vaccinated. He obtained the following figures: never, 12 people; once, 24; twice, 42; three times, 38; four times, 30; five times, 4. Based on this data:

- What is the mean number of times those people had been vaccinated and what is the standard deviation?
- Is the standard deviation a good measure of variation in this case?

8. From the 140 children whose urinary concentration of lead was investigated, 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper (in $\mu\text{mol}/24\text{h}$) were found.

0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77, 0.65, 0.76, 0.42, 0.94, 0.36, 0.98, 0.64, 0.90, 0.63, 0.55, 0.78, 0.10, 0.52, 0.42, 0.58, 0.62, 1.12, 0.86, 0.74, 1.04, 0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88.

- Find the mean, median, mode, range, and quartiles.

- Play with the data and change the value 1.24 to 2.24. See how the statistics found in the last exercise are affected.

- With the mean you obtained in the first exercise, obtain the standard deviation, and an approximate 95% range. Which points are excluded from the range mean -2SD to mean $+2\text{SD}$?

9. The following data were found in a study of asthmatic children. What are the best ways of graphically displaying the summaries of these data?

- Peak flow: quantitative data and symmetrically distributed.
- Number of episodes of wheeziness per day: quantitative data with a skewed distribution.
- Social class of the child's parents: qualitative and categorical data.

10. The mean urinary lead concentration in 140 children was $2.18\mu\text{mol}/24\text{h}$ with standard deviation 0.87.

- What is the standard error of the mean?

11. Multiple-choice questions.

1. Find the median of the set of numbers: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.

- 1.
- 5.5.
- 10.
- 55.
- 100.

2. The summation of the linear deviations from the mean for a set of data will always be:

- A positive number.
- A negative number.
- Zero.
- The absolute value of the mean.
- None of the above.

3. Which of the following is not a measure of the variability of data?

- a) Variance.
- b) Range.
- c) Mean.
- d) Standard deviation.
- e) None of the above.

4. Which one of these statistics is unaffected by outliers?

- a) Mean.
- b) Interquartile range.
- c) Standard deviation.
- d) Range

5. Which of the following would indicate that a dataset is not bell-shaped?

- a) The range is equal to 5 standard deviations.
- b) The range is larger than the interquartile range.
- c) The mean is much smaller than the median.
- d) There are no outliers.

6. The following are the quartiles and median for the times of 100 swimmers aged 19–29 in the Newport one-mile swim: Q1 = 24 minutes, median = 26 minutes and 45 seconds, Q3 = 28 minutes and 21 seconds. About what percent of the swimmers had times in the interval from 24 minutes to 28 minutes and 21 seconds?

- a) 25%.
- b) 50%.
- c) 75%.
- d) 100%.

7. Which statement is not true about confidence intervals?

- a) A confidence interval is an interval of values computed from sample data that is likely to include the true population value.
- b) An approximate formula for a 95% confidence interval is sample estimate \pm margin of error.
- c) A confidence interval between 20% and 40% means that the population proportion lies between 20% and 40%.
- d) A 99% confidence interval procedure has a higher probability of producing intervals that will include the population parameter than a 95% confidence interval procedure.

8. Based on the empirical rule for the normal distribution, about what percentage of the values are within two standard deviations of the mean.

- a) 50%.
- b) 60%.
- c) 68%.
- d) 95%.
- e) 99%.

9. What shape is a normal distribution?

- a) Bi-modal.
- b) Inverted (U).
- c) Bell shaped.
- d) Ascending line.
- e) Descending line.

10. Which measure is the most unreliable indicator of central tendency if data are skewed?

- a) Distribution.
- b) Median.
- c) Range.
- d) Mode.
- e) Mean.

11. What type of distribution is observed when most of the scores cluster around the higher end of the scale?

- a) A positively skewed distribution.
- b) A related distribution.
- c) A negatively skewed distribution.
- d) A normal distribution.
- e) A bi-modal distribution.

12. What type of distribution is observed when most of the scores cluster around the lower end of the scale?

- a) A positively skewed distribution.
- b) A related distribution.
- c) A negatively skewed distribution.
- d) A normal distribution.
- e) A bi-modal distribution.

13. Which two measures use the mean as a baseline and identify the extent to which scores differ from this?

- a) Variance and standard deviation.
- b) Standard deviation and median.
- c) Mode and median.
- d) Standard deviation and range.
- e) Sum and variance.

14. The Central Limit Theorem says that the mean of the sampling distribution of the sample means is:

- a) Equal to the population mean divided by the square root of the sample size.
- b) Close to the population mean if the sample size is large.
- c) Exactly equal to the population mean.
- d) None of the above.

15. Which one of the following statements is true?

- a) A pie chart is one in which a circular 'pie' is split into sectors, one for each category of a categorical variable, so that the area of each sector is equal.
- b) A sensible way of displaying continuous numerical data is to draw a bar chart.
- c) A histogram is a chart in which separate vertical (or horizontal) bars are drawn with gaps between the bars; the width (height) of each bar relates to a specific range of values of the variable, and its height (width) is proportional to the associated frequency of observations.
- d) The distribution of a variable is right skewed if a histogram of observed values has a long tail to the right with one or a few high values.
- e) A box-and-whisker plot comprises a vertical or horizontal rectangle indicating the interquartile range, within which is the median; the ends of the 'whiskers' represent the upper and lower limits of the 95% confidence interval for the median.

16. A 95% confidence interval is:

- a) The range in which a mean value falls approximately 95% of the time.
- b) The range in which 95% of the study observations can be expected to lie.
- c) The range in which we are 95% certain that the true population value lies.
- d) The range calculated as the mean \pm 1.96 standard deviations and which excludes 5% of the sample.

17. If a set of observations follow the Normal or distribution, which one of the following statements is true?

- a) Its mean and variance are equal.
- b) Its observations are derived from healthy individuals.
- c) Its mean and variance are always equal to zero and one, respectively.
- d) 95% of the observations lie between the mean ± 1.96 times the variance.
- e) Approximately 68% of the observations lie between the mean \pm the standard deviation.

18. When numerical data are arranged in order of magnitude, which one of the following statements is true?

- a) The interquartile range is the difference between the first and fourth percentiles.
- b) The interdecile range contains the central 80% of the ordered observations.
- c) The middle observation is always equal to the arithmetic mean.
- d) The 50th percentile is equal to the fifth quartile.
- e) The first percentile is always equal to the minimum value.

19. Which one of the following statements is true?

- a) The median is greater than the arithmetic mean if the data are skewed to the right.
- b) The median value of n observations is equal to the $(n + 1)/2$ th value in the ordered set if n is odd.
- c) The median and the weighted mean are always identical if the weights used in the calculation of the weighted mean are equal.
- d) The logarithmic transformation of left-skewed data will often produce a symmetrical distribution when the transformed data are plotted on an arithmetic scale.
- e) The geometric mean of a data set is equal to the arithmetic mean of the log-transformed data.

20. As part of an epidemiological study investigating the association between consumption of dairy products in adolescence and the onset of cardiovascular disease later in life, study investigators plan to collect information on weekly egg consumption from a sample of children aged 14–17 years using self-administered questionnaires. The authors wish to summarize the data on the number of eggs consumed in a week. Which one of the following approaches would be the best way to summarize these data?

- a) The arithmetic mean and range.
- b) The median and interquartile range.
- c) The median and range.
- d) The arithmetic mean and standard deviation.
- e) The mode.

Bibliography and Suggested Reading

- Bowers D. Medical Statistics from Scratch. An Introduction for Health Professionals. 3rd Edition. West Sussex: John Wiley & Sons, Inc; 2014.
- Collaco JM, Green DM, Cutting GR, Naughton KM, Mogayzel PJ Jr. Location and duration of treatment of cystic fibrosis respiratory exacerbations do not affect outcomes. *Am J Respir Crit Care Med.* 2010; 182: 1137–43.
- Carlson KA, Winquist JR. An Introduction to Statistics. An Active Learning Approach. 2nd Edition. Thousand Oaks: SAGE Publications; 2018.
- Dtsch Arztebl Int 2009; 106(36): 578–83 DOI: 10.3238/arztebl.2009.0578.
- Esparza-Villalpando V. Basic Biostatistics with R. MCIC. San Luis Potosí; 2019.
- Kestin I. Statistics in medicine. *Anaesthesia and Intensive Care Medicine.* 2012;13(4):200–207.
- Machin D, Campbell MJ, Walters SJ. Medical Statistics. A Textbook for the Health Sciences. 4th Edition. West Sussex: John Wiley & Sons; 2007.
- Pezzullo JC. Biostatistics for Dummies. Hoboken, NJ: John Wiley & Sons, Inc; 2013.

Inferential Statistics

Learning objectives for this chapter

- A. Understand hypothesis testing as making an argument.
- B. Identify the steps of hypothesis testing.
- C. Define null hypothesis, alternative hypothesis, level of significance, test statistic, p-value, and statistical significance.
- D. Recognize that the strength of evidence against the null hypothesis depends on how unlikely it would be to get a statistic as extreme just by random chance, if the null hypothesis were true.
- E. Demonstrate an understanding of the concept of statistical significance.
- F. Interpret the p-value.
- G. Recognize Type I and Type II error, and interpret them in context.

In Inferential Statistics, the main aim is to use the information obtained from a sample of individuals to **make inferences about the population of interest**. It is never known how representative this sample is, so these inferences are always made with some **uncertainty**. This uncertainty is measured by **probabilities**, and these probabilities measure the **degree of confidence of our conclusions** about the population.

There are two basic approaches to statistical analysis: **Estimation** (with Confidence Intervals) and **Hypothesis Testing** (with p-values).

Hypothesis

All hypothesis tests are based on specific **assumptions**. If these assumptions are violated, these tests may yield **misleading results**. Therefore, the first step when conducting any hypothesis test is to determine if the data you are about to analyze meet the **four basic assumptions**.

Basic Assumptions for Hypothesis Testing

- 1. Independence of the data:** Each participant's score within a condition is independent of all other participants' scores within that same condition.
- 2. Appropriate measurement of variables for the analysis:** The independent variable must identify a group of people who are different from the population in some way, and the dependent variable must be measured on an interval or ratio scale of measurement.
- 3. Normality of distributions:** The distribution of sample means for each condition must have a normal shape.
- 4. Homogeneity of variance:** The variances in each condition of the study are similar.

Hypothesis Testing

Statistical analysis is concerned not only with summarizing data but also with **investigating relationships**. If an investigator conducting a study has a theory in mind, his theory is known as the **study or research hypothesis**. However, it is impossible to prove most hypotheses; one can always think of circumstances which have not yet arisen under which a particular hypothesis may or may not hold. Thus, there is a **simpler logical setting for disproving hypotheses** than for proving them.

The opposite of the research hypothesis is the **null hypothesis** (H_0 , read as "H-naught"). Such hypothesis is usually **phrased in the negative** and that is why it is termed null. The null hypothesis predicts that the difference observed "**occurred by chance**".

Unfortunately, **only one of the hypothesis can be correct**.

Table 10.1 states a representation of the hypothesis based on the following example:

A teacher predicts that frequent quizzing will increase test scores in his students.

- » His research hypothesis states that the student who take frequent quizzes will have a higher mean than the population of students who did not take frequent quizzes.
- » The null hypothesis is that frequent quizzing will either have no effect on test scores or will decrease them.

Table 10.1. Research and Null Hypothesis

Hypothesis Type	Symbol	Verbal	The Difference Between Sample and Population Means was Created by
Research hypothesis	$\mu_{\text{Quiz}} > 7.5$	The population of people who take frequent quizzes will have quiz scores higher than 7.5	The treatment improving final exam scores
Null hypothesis	$\mu_{\text{Quiz}} \leq 7.5$	The population of people who take frequent quizzes will not have quiz scores higher than 7.5	Sampling error

Since we've stated before that only one hypothesis can be correct, the entire point of performing hypothesis testing is to determine **which of these two hypothesis is most likely to be true**. To do this, we strongly recommend to follow the following critical path:

1. It is assumed that **the parameters of both populations are identical and there is no difference between them** (this defines our null hypothesis, H_0).
2. The **sample statistics are calculated**. If the sample statistics calculated are identical, it could be concluded that there are no differences between the populations, and the study would be terminated.
3. If the statistics calculated from the samples are different, in principle you could think that the difference is **because of chance**. In other words, since until now the null hypothesis is considered to be true, the difference is explained because in one of the populations, elements with different values from those selected in the other population have been selected by chance.
4. The following question is immediately asked: if the samples that are being compared come from the same universe, how likely is it to observe a difference like the one being observed? To answer it, we'll have to **calculate the probability that the difference occurred by chance**.

5. Finally, you have to pick one of the following options:

- If there is a **high probability** that the results obtained **occurred by chance**, then it cannot be ruled out that **chance** is the explanation of the difference observed and **the null hypothesis is not rejected**.
- If there is a **low probability** that the results observed are **due to chance**, then it is thought that **there must be another explanation** for the differences found and **the null hypothesis is rejected**.

“**Not rejecting a null hypothesis**” means that we do not find evidence that allows us to assume that the two populations are different, since the difference found may, with great probability, be due to chance.

» **Caution:** When we do not reject a null hypothesis, we generally avoid affirming that the populations studied are equal.

Rejecting a null hypothesis implies that the difference observed is unlikely to be explained by chance, so **the difference must be explained by another cause**. Consequently we suggest a second hypothesis that we call an **alternate hypothesis, H1**.

Most often, we try to test our research hypothesis by testing the alternate hypothesis. But **beware**, the only thing that proves the alternative hypothesis is that chance is an unlikely cause as an explanation of the differences found.

To **verify a statistical hypothesis** you can proceed according to the following sequence:

1. Statement of the hypothesis: State a null hypothesis (statement you are looking for evidence to disprove), and an alternative hypothesis (HA).
2. Selection of the level of significance: The α of the test.
3. Description of the population of interest and approach to the necessary assumptions.
4. Selection of the relevant statistic.
5. Specification of the test statistic and consideration of its distribution.
6. Specification of rejection and acceptance regions.
7. Data collection and calculation of the necessary statistics.
8. Statistical decision.
9. Conclusion.

Statistical Significance

All the famous statistical significance tests (Student t test, chi-square, ANOVA, etc.) work on the same general principle: **they evaluate the size of the apparent effect you see in your data against the size of the random fluctuations present in your data.**

In order to test for significance, it's useful to follow this general two steps:

1. Obtain a test statistic.

- Each test has its own formula, with which we will obtain a final result (**test statistic**).
- This test statistic represents the **magnitude of the effect you're looking** for in relation with the magnitude of the random noise in your data.

2. Determine how likely it is for random fluctuations to produce a test statistic as large as the one obtained from the data.

- There are formulas that describe how much the test statistic bounces around if only random fluctuations are present (that is, if H_0 is true).

Criterion for Significance

We take the value α as the cut-off point to determine what results we are going to consider as “statistically significant”. Usually, the value for α is set as **0.05** (more on this in the Type I error section).

Useful tip:
When the sample size is big enough, even small differences observed can be considered as “statistically significant”.

The p-value

The medical journals are replete with p-values and tests of hypotheses. It is a common practice among medical researchers to quote whether the test of hypothesis they carried out is significant or non-significant, and many researchers get very excited when they discover a “statistically significant” finding without really understanding what it means.

Where Does the p-value Came From?

The prominence of the p-value in the scientific literature is attributed to **R. A. Fisher**, who did not invent this probability measure but did popularize its extensive use for all forms of statistical research methods starting with his seminal 1925 book, “Statistical Methods for Research Workers”.

According to Fisher, the correct definition of the p-value is:

- » “The probability of the observed result, plus more extreme results, if the null hypothesis were true.”

Fisher’s purpose was not to use the p-value as a decision-making instrument but to provide researchers with a **flexible measure of statistical inference** within the complex process of scientific inference. In addition, there are important assumptions associated with proper use of the p-value.

How to Interpret a p-value?

The p-value is used in all statistical tests, from t tests to regression analysis. Despite being so important, the p-value is a difficult concept and many often interpret it incorrectly.

The p-value should be understood as the **proportion of times that the contrast statistic** (mean, standard deviation, variance, proportion, etc.) **takes a more extreme** (different) **value than the result of the experiment performed**.

Fisher developed the criterion for “statistically significant” when the p-value is **lower than 0.05**. This criterion of **95% confidence** (or 0.05 probability), is the **basis of modern Statistics**.

The significance testing that we use today is based on the Fisher’s idea using the p-value as an **index** of the weight of **evidence against a null hypothesis**. The cut-off point of p-value lower than 0.05 is **not totally explained**, however, the use of dogmatic and fixed cut-off point does not work in all circumstances. The use of **effect sizes** regulate the balance between the use of arbitrary cut-off points and **indicate the meaningful effect of the difference**.

Using p-values to Make a Decision About Whether to Reject or Not Reject the Null Hypothesis

Based on the premise that the results are statistically significant if the p-value is less than 0.05, whenever you are confronted with determining whether or not you should reject the null hypothesis, you can use either one of the following two rules:

- » If the obtained value is **more extreme** than the critical value, you should **reject the null hypothesis**.
- » If the p value is **less than the α value**, you should **reject the null hypothesis**.

The **p-value** can be understood as the probability to find a value of the statistic of contrast farther or more extreme than what was observed in the current sample if we repeat the experiment in the same conditions in an infinitely manner.

Statistical Power: Type I and Type II Errors

Type I Error

If you **reject the null hypothesis when it is in fact true**, then you'll be making a Type I error (**false-positive**).

Statisticians use the Greek letter alpha (α) to represent the probability of making a Type I error. The quantity α is interchangeably termed the **test size**.

Limiting your chance of making a Type I error (falsely claiming significance) is very easy. If you don't want to make a Type I error more than 5 percent of the time, **don't declare significance unless the p-value is less than 0.05**. That's called testing at the 0.05 α level.

- » If you're willing to make a Type I error 10 percent of the time, use $p < 0.10$ as your criterion for significance.

Useful tip:
Type I error means that there are statistically significant differences when actually there are not.

Type II Error

If you **do not reject the null hypothesis when it is in fact not true**, then you'll be making a Type II error (**false-negative**).

Statisticians use the Greek letter beta (β) to represent the probability of making a Type II error. The **power of the study** is defined as one minus the probability of a Type II error, thus the power equals $1 - \beta$. That is, the **power** is the probability of obtaining a "statistically significant" p-value when the null hypothesis is truly false.

Usually, the value for β is set as 0.10–0.20.

Limiting your chance of making a Type II error (falsely not claiming significance) is not quite easy. If you don't want to make a Type II error, be sure to have a **sufficient sample size** in your study.

Useful tip:
Type II error means that there are not statistically significant differences when actually there are.

The relationship between Type I and II errors and significance tests is shown in **Table 10.2**, and is illustrated in **Figure 10.1**.

The concepts of Type I error and Type II error parallel the concepts of sensitivity and specificity (**Chapter 22**).

- » The Type I error is equivalent to the **false positive rate (FPR)**.
 - $FPR = 1 - \text{specificity}$.
- » The Type II error is equivalent to the **false negative rate (FNR)**.
 - $FNR = 1 - \text{sensitivity}$.

Bridge to Diagnostic Tests

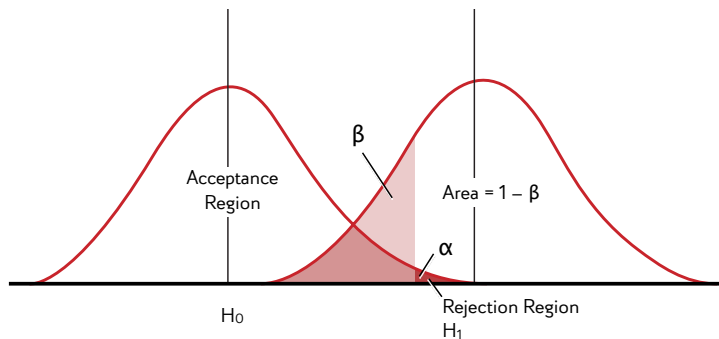


Sensitivity: Proportion of true positive (diseased) with a positive test result.

Specificity: Proportion of true negative (non-diseased) with a negative test result.

Table 10.2. Research and Null Hypothesis

The Truth (based on the Entire Population)		
Your Conclusion (based on the sample)	Difference Does Not Exist (H_0 is true)	Difference Does Exist (H_0 is false)
Non-significant	Right! That's a true negative case	Type II error (β)
Significant	Type I error (α)	Right! That's a true positive case

**Figure 10.1.** Theoretical visual representation of acceptance and rejection regions, alpha (α) and beta (β) error regions, and power.

Key Terms

Define the following terms.

Alpha (α)

Alternate hypothesis

Appropriate measurement of variables

Assumption

Beta (β)

Confidence

Estimation

False-negative

False-positive

Homogeneity of variance

Hypothesis

Hypothesis testing

Independence of the data

Inferential statistics

Normality of distributions

Null hypothesis

p-value

Probability

Rejecting hypothesis

Research hypothesis

Statistical power

Statistical significance

Test statistic

Type I error

Type II error

Uncertainty

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Answer the following questions:

- Can 95% confidence intervals be used to infer P values and vice versa?
- When might a significance test fail to detect a real effect?
- When is the null hypothesis value outside the 95% confidence interval?
- What type of error occurs when a difference between groups is not statistically significant but is large enough to be thought clinically important?
- Who decides what size of difference between groups is clinically important?

2. A clinical trial to compare a mouthwash against a control found a difference in plaque score after 1 year of 1.1 units, $P = 0.006$ (two sided). State if the following are true or false:

- The probability that the null hypothesis is true is 0.006.
- If the null hypothesis were true, the probability of getting an observed result of 1.1 or greater is 0.003.
- The alternative hypothesis is a mean difference of 1.1.
- The probability of the alternative hypothesis being true is 0.994.
- The probability that the true mean is 1.1 units is 95%.

3. Multiple choice questions.

1. Hypothesis testing procedures were created so that researchers could:

- Study entire populations rather than samples.
- Deal with sampling error.

2. Testing casual hypotheses requires knowing how to:

- Use statistics.
- Use research methods to design “fair” experiments.
- Both of the above.

3. To conduct a hypothesis test using the z for a sample mean, the dependent variable must be measured on a:

- Ordinal scale.
- Interval/ratio scale.
- Either ordinal or interval/ratio scale.

4. The fact that the null and research hypotheses are mutually exclusive means that if the null is true:

- The research hypothesis must also be true.
- The research hypothesis can be true or false.
- The research hypothesis must be false.

5. Even if the null hypothesis is true, you should not be surprised if the z score resulting from a significance test is not the exact value you selected above because of _____.

- A nonspecific research hypothesis.
- Sampling error.

6. If the null hypothesis is true, z scores close to zero are:

- Likely.
- Unlikely.

7. When the research hypotheses is that the sample mean will be lower than the population mean, the critical region will be on the _____ side of the distribution because they are expecting that the obtained z score will be _____.

- a) Positive, positive.
- b) Negative, positive.
- c) Negative, negative.

8. Which one of the following statements is true?

- a) If the hypothesized value for the effect of interest (e.g. the difference in means) in a hypothesis test lies within the 95% confidence interval for the effect, then we have evidence to reject the hypothesis, $p < 0.05$.
- b) A hypothesis test of superiority which proceeds by calculating a test statistic and relating it to the appropriate probability distribution to obtain the p-value is so called because is it superior to testing the hypothesis using the relevant confidence interval.
- c) The test statistic that is calculated in a hypothesis testing procedure reflects the amount of evidence in the data against the null hypothesis.
- d) A bioequivalence trial is a particular type of randomized trial which is concerned with demonstrating that biological treatments have the same effect as non-biological treatments on a disease outcome.
- e) Nonparametric tests lead to an appreciation of the data, rather than focusing on decisions, because they do not concentrate on the parameters of the underlying distributions.

9. Which one of the following statements is true about the probability of making a Type I error when performing a single hypothesis test?

- a) It is equal to one minus the probability of a Type II error.
- b) It is the probability of rejecting the null hypothesis when it is true.
- c) It is the probability of not rejecting the null hypothesis when it is false.
- d) It can never exceed 0.05.
- e) It is equal to the significance level of the hypothesis test.

10. A type II error occurs when:

- a) A statistician makes an error in calculating a p-value.
- b) An important difference between groups has a p-value larger than 0.05.
- c) A clinically important effect is unlikely to have occurred by chance.
- d) A new treatment proves more effective than was thought when the sample size was calculated.

11. In hypothesis testing, a Type II error occurs when:

- a) The null hypothesis is not rejected when the null hypothesis is true.
- b) The null hypothesis is rejected when the null hypothesis is true.
- c) The null hypothesis is not rejected when the alternative hypothesis is true.
- d) The null hypothesis is rejected when the alternative hypothesis is true.

12. A hypothesis test is done in which the alternative hypothesis is that more than 10% of a population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

- a) We can conclude that more than 10% of the population is left-handed.
- b) We can conclude that more than 25% of the population is left-handed.
- c) We can conclude that exactly 25% of the population is left-handed.
- d) We cannot conclude that more than 10% of the population is left-handed.

13. A result is called “statistically significant” whenever:

- a) The null hypothesis is true.
- b) The alternative hypothesis is true.
- c) The p-value is less or equal to the significance level.
- d) The p-value is larger than the significance level.

14. A test to screen for a serious but curable disease is similar to hypothesis testing, with a null hypothesis of no disease, and an alternative hypothesis of disease. If the null hypothesis is rejected, treatment will be given. Otherwise, it will not. Assuming the treatment does not have serious side effects, in this scenario it is better to increase the probability of:

- a) Making a Type 1 error, providing treatment when it is not needed.
- b) Making a Type 1 error, not providing treatment when it is needed.
- c) Making a Type 2 error, providing treatment when it is not needed.
- d) Making a Type 2 error, not providing treatment when it is needed.

15. A significance test based on a small sample may not produce a statistically significant results even if the true value differs substantially from the null value. This type of result is known as:

- a) The significance level of the test.
- b) The power of the study.
- c) A type I error.
- d) A type II error.

16. One problem with hypothesis testing is that a real effect may not be detected. This problem is most likely to occur when:

- a) The effect is small and the sample size is small.
- b) The effect is large and the sample size is small.
- c) The effect is small and the sample size is large.
- d) The effect is large and the sample size is large.

17. The probability of Type I error is referred as?

- a) $1-\alpha$.
- b) β .
- c) α .
- d) $1-\beta$.

18. Which of the following statements are true?

- a) The p-value is the probability of the sample data arising by chance.
- b) The p-value is an arbitrary value, designated as the significance level.
- c) The p-value is the chance of getting an observed effect if the null hypothesis was false.
- d) The p-value is the chance of getting an observed effect if the null hypothesis was true.
- e) A very small p-value allows us to say that there is enough evidence to accept the null hypothesis.

19. The result of a statistical test, denoted p , shall be interpreted as follows:

- a) The null hypothesis H_0 is rejected if $p < 0.05$.
- b) The null hypothesis H_0 is rejected if $p > 0.05$.
- c) The alternate hypothesis H_1 is rejected if $p > 0.05$.
- d) The null hypothesis H_0 is accepted if $p < 0.05$

20. A researcher believes that the proportion of individuals with diagnosed epilepsy that present with a depressive disorder is higher than the proportion of individuals without diagnosed epilepsy that present with a depressive disorder. Testing this claim, the resulting p -value is 0.003. Using a 0.10 significance level, which of the following is the most appropriate conclusion given the results?

- a) Reject the null hypothesis; there is sufficient evidence to support the researcher's claim.
- b) Fail to reject the null hypothesis; there is sufficient evidence to support the researcher's claim.
- c) Accept the null hypothesis; there is not sufficient evidence to support the researcher's claim.
- d) Accept the null hypothesis; there is sufficient evidence to support the researcher's claim.

Bibliography and Suggested Reading

- Bowers D. Medical Statistics from Scratch. An Introduction for Health Professionals. 3rd Edition. West Sussex: John Wiley & Sons, Inc; 2014.
- Carlson KA, Winquist JR. An Introduction to Statistics. An Active Learning Approach. 2nd Edition. Thousand Oaks: SAGE Publications; 2018.
- Dahiru T. p -value, a True Test of Statistical significance? A Cautionary Note. Annals of Ibadan Postgraduate Medicine. 2008;6(1):21-26.
- Dtsch Arztebl Int 2009; 106(19): 335–9 DOI: 10.3238/arztebl.2009.0335.
- Esparza-Villalpando V. Basic Biostatistics with R. MCIC. San Luis Potosí; 2019.
- Kain ZN et al. Valor de p inferior a 0,05: ¿qué significa en realidad? Pediatrics (Ed esp). 2007;63(3):118-20.
- Kyriacou DN. The Enduring Evolution of the P Value. JAMA. 2016;315(11):1113-1115.
- Machin D, Campbell MJ, Walters SJ. Medical Statistics. A Textbook for the Health Sciences. 4th Edition. West Sussex: John Wiley & Sons; 2007.
- Molina Arias M. ¿Qué significa realmente el valor de p ? Rev Pediatr Aten Primaria. 2017;19:377-81.
- Pezzullo JC. Biostatistics for Dummies. Hoboken, NJ: John Wiley & Sons, Inc; 2013.

Statistical Tests

Learning objectives for this chapter

- A. Recognize when and why statistical tests are needed.
- B. Identify appropriate statistical methods to be applied in a given research setting, and acknowledge the limitations of those methods.
- C. Know the fundamentals of the most relevant parametric and nonparametric techniques for statistical inference.
- D. Choose and set up suitable parametric methods for hypothesis testing estimation.
- E. When appropriate, choose and set up suitable non-parametric methods for hypothesis testing estimation.
- F. Distinguish between non-parametric and parametric tests.
- G. Explain why ordinal data are computed using nonparametric tests.

How to Choose a Statistical Test?

In order to have an appropriate approach to the statistical analysis of the collected data, you can ask yourself this five questions:

1. What are the aims and objectives of the study?
2. What is the hypothesis to be tested?
3. What type of data are the outcome data?
4. How is the outcome data distributed?
5. What is the summary measure for the outcome data?

Although it may seem a little too much repetitive, **Table 11.1** and **Table 11.2** list the most important statistical tests used. Each Table deals with them in different ways, trying to facilitate and promote their better understanding.

In case you are a student that performs better with text, have no fear, the statistical tests are **concisely described** below.

Chi-square:

Tests for the strength of the association between two categorical variables.

Wilcoxon rank-sum test:

Tests for difference between two independent variables - takes into account magnitude and direction of difference.

Wilcoxon sign-rank test:

Tests for difference between two related variables - takes into account magnitude and direction of difference.

Table 11.1. Frequently Used Statistical Tests

Statistical Test	Description
Fisher's exact test	Suitable for binary data in unpaired samples: the 2 x 2 table is used to compare treatment effects or the frequencies of side effects in two treatment groups
Chi-square test	Similar to Fisher's exact test (albeit less precise) Can also compare more than two groups or more than two categories of the outcome variable Preconditions: sample size >60, expected number in each field ≥5
McNemar test	Preconditions similar to those for Fisher's exact test, but for paired samples
Student's t-test	Test for continuous data . Investigates whether the expected values for two groups are the same, assuming that the data are normally distributed. The test can be used for paired or unpaired groups
Analysis of Variance	Test preconditions as for the unpaired t-test, for comparison of more than two groups. The methods of analysis of variance are also used to compare more than two paired groups
Wilcoxon's rank sum test (also known as the unpaired Wilcoxon rank sum test or Mann-Whitney U test)	Test for ordinal or continuous data . In contrast to Student's t-test, does not require the data to be normally distributed. This test can also be used for paired or unpaired data
Kruskal-Wallis test	Test preconditions as for the unpaired Wilcoxon rank sum test for comparing more than two groups
Friedman test	Comparison of more than two paired samples, at least ordinally scaled data
Log rank test	Test of survival time analysis to compare two or more independent groups
Pearson correlation test	Tests whether two continuous normally distributed variables exhibit linear correlation
Spearman correlation test	Tests whether there is a monotonous relationship between two continuous , or at least ordinal , variables

Table 11.2. Statistical Tests Based on the Type of Data

Type of Data	Two Groups	More Than Two Groups
Categorical data (e.g. blood group)	Contingency tables	Contingency tables
Ordinal data (e.g. Glasgow coma scale)	Unpaired: Mann-Whitney test Paired: Wilcoxon Rank Sum test	Unpaired: Kruskal-Wallis test Paired: Friedman's test
Continuously variable data, normally distributed (e.g. weight)	Unpaired: t-test Paired: paired t-test	Unpaired: ANOVA Paired: paired ANOVA
Continuously variable data, not normally distributed (e.g. duration of hospital stay)	As for ordinal data Or transform the data into a normal distribution	As for ordinal data Or transform the data into a normal distribution

In some studies, each subject in one group may be uniquely paired with one in the other group(s). For example, if a variable is measured in the same individual before and after an intervention, then these two observations are “**paired**”. There are appropriate statistical tests that should be used for this type of data.

Common Problems in Statistical Inference

When trying to understand statistical inference, you may find two scenarios:

- » Comparison of **independent groups** (e.g., groups of patients given different treatments).
- » Comparison of the response for **paired observations** (e.g., in a cross-over trial or for matched pairs of subjects).

In order to simplify the analysis of these data, two algorithms are proposed: **Figure 11.1** for comparing independent groups and **Figure 11.2** for comparing paired samples. Furthermore, **Table 11.3** contains the names of several statistical procedures classified as **parametric** or **nonparametric**.

- » **Parametric tests:** assume a normal distribution of values, or a “bell-shaped curve.”
- » **Nonparametric tests:** used in cases where parametric tests are not appropriate.

Paired data:

Occur when natural matching or coupling is possible. Generally this would be data sets where every data point in one independent sample would be paired—uniquely—to a data point in another independent sample. Paired data must be analyzed as such; and cannot be dealt as independent data.

Some common situations for using nonparametric tests are:

- When the distribution is not normal (the distribution is skewed).
- When the distribution is not known.
- When the sample size is too small to assume a normal distribution.
- If there are extreme values or values that are clearly out of range.

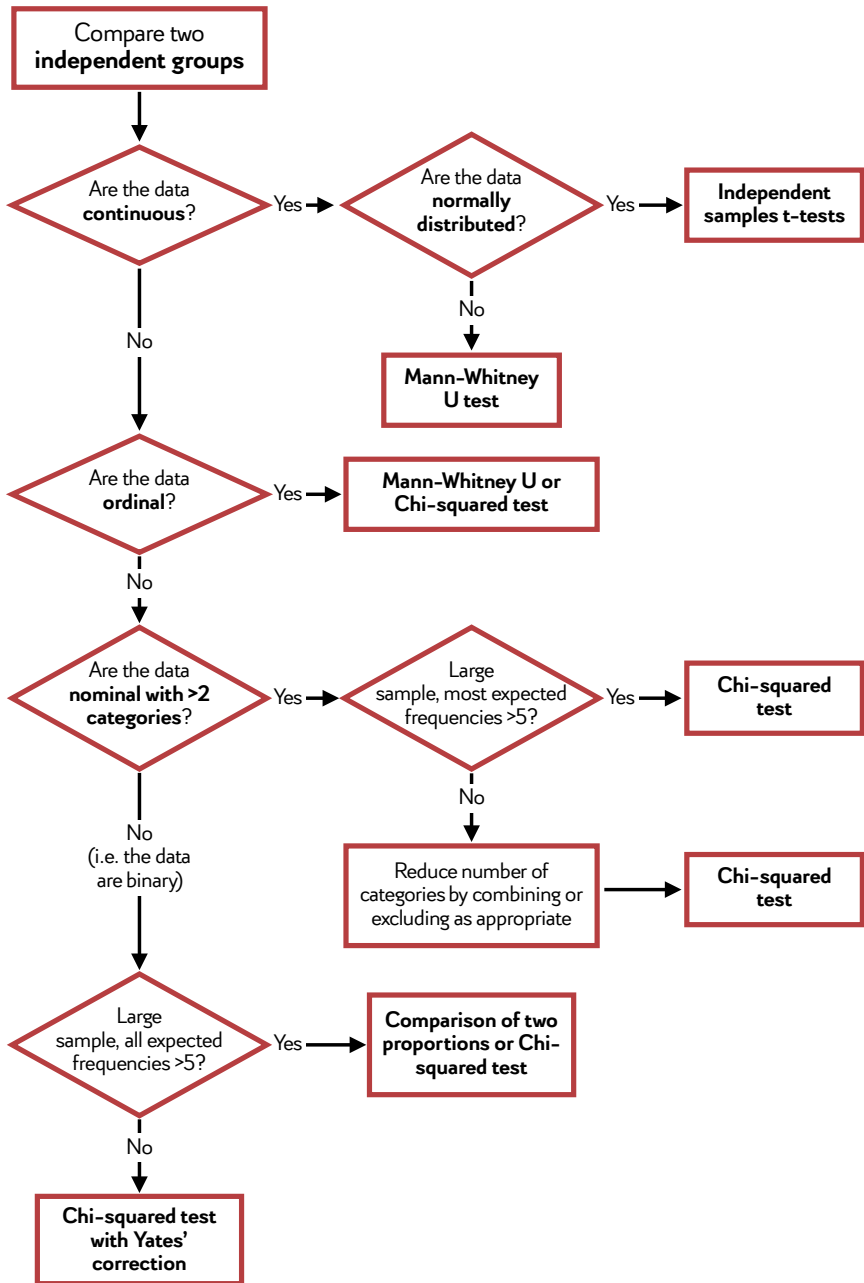


Figure 11.1. Algorithm for comparing independent groups of data.

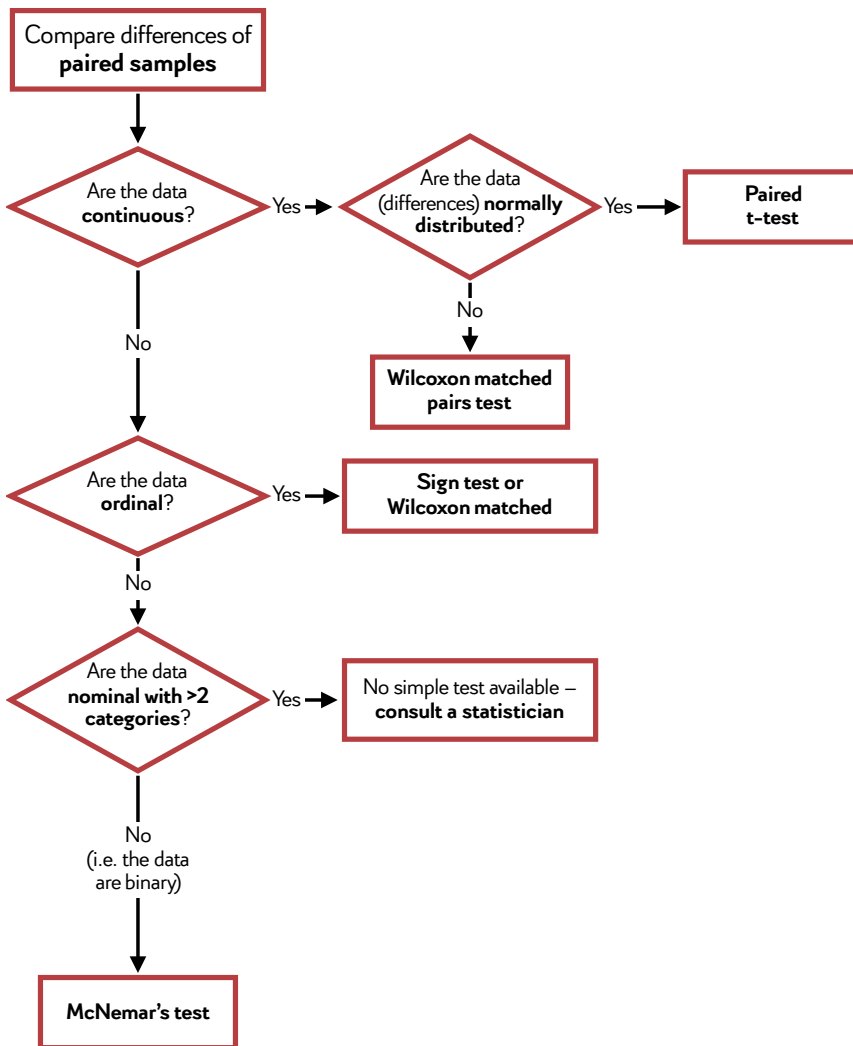


Figure 11.2. Algorithm for comparing paired samples of data.

Table 11.3. Parametric and Nonparametric Tests

Analysis Type	Example	Parametric Test	Nonparametric Test
Compare means between two independent groups	Is the mean systolic blood pressure (at baseline) for patients assigned to placebo different from the mean for patients assigned to the treatment group?	Student's t-test	Wilcoxon rank-sum test
Compare two quantitative measurements taken from the same individual	Was there a significant change in systolic blood pressure between baseline and the six-month follow-up measurement in the treatment group?	Paired t-test	Wilcoxon rank-sum test
Compare means between three or more distinct/independent groups	If our experiment had three groups (e.g., placebo, new drug #1, new drug #2), we might want to know whether the mean systolic blood pressure at baseline differed among the three groups?	Analysis of variance (ANOVA)	Kruskal-Wallis test
Estimate the degree of association between two quantitative variables	Is systolic blood pressure associated with the patient's age?	Pearson coefficient of correlation	Spearman's rank correlation

Comparing Average Values Between Groups

Is part of the analysis of almost every biological experiment, therefore there are dozens of tests for this purpose. These tests include different flavours of the Student's t-test, as well as:

- » Analyses of variance (ANOVA).
- » Analyses of covariance (ANCOVA).
- » Wilcoxon.
- » Mann-Whitney.
- » Kruskal-Wallis.
- » Friedman.

One thing must be clear: Different situations need different tests, because “comparing averages” does not refer to a single task, rather it applies to a lot of situations that differ based on:

- » Whether you're looking at changes over time within one group of subjects or differences between groups of subjects (or both).
- » How many time points or groups of subjects you're comparing.
- » Whether or not the numeric variable you're comparing is nearly normally distributed.
- » Whether or not the numbers have the same spread (standard deviation) in all the groups you're comparing.
- » Whether you want to compensate for the possible effects of some other variable on the variable you're comparing.

Comparing the Mean of a Group of Numbers to a Hypothesized Value

Some studies do not have a control group. Therefore, the results must be compared to a “historical” control (available in the information from the literature).

- » This data are usually analyzed by the **one-group Student t-test**.
- » If the data are non-normal, the **Wilcoxon Signed-Ranks test** can be used.

Comparing Two Groups of Numbers

Is one of the most common situations in clinical research. these comparisons are handled by the **unpaired** or “**independent sample**” **Student t-test** (or merely “t-test”).

The variance is simply the square of the standard deviation.

This **t-test** is based on two assumptions:

- » **Normality assumption:** The data are normally distributed.
 - For non-normal data you can use the nonparametric **Mann-Whitney (M-W) test**.
- » **Equal-variance assumption:** The standard deviation (SD) is the same for both groups.
 - If the two groups have noticeably different variances, then the t-test may not give reliable results. Instead, you can use a modification to the Student t-test, called the **Welch test** (also called the Welch t-test, or the unequal-variance t test).

Comparing Three or More Groups of Numbers

This can be assumed as an extension of the two-group comparison, and is handled by the **analysis of variance (ANOVA)**.

The ANOVA is a very general method that can accommodate several grouping variables at once. For example, suppose you want to compare the response to treatment among three treatment groups (i.e., drug A, drug B, and placebo).

The term **way** refers to how many grouping variables are involved

- » When there is **one grouping variable** (e.g., treatment), you have a **one-way ANOVA**.
- » An ANOVA involving **two different grouping variables** is called a **two-way ANOVA**.
- » An ANOVA involving **three different grouping variables** is called a **three-way ANOVA**.

The **null hypothesis of the one-way ANOVA** is that all the groups have the same mean.

The **alternative hypothesis of the one-way ANOVA** is that at least one group is different from at least one other group.

Like the t-test, the ANOVA also assumes normally distributed numbers with equal standard deviations in all the groups.

- » If your data is non-normal, you can use the **Kruskal-Wallis test** instead of the one-way ANOVA.
- » If the groups have very dissimilar standard deviations, you can use the **Welch unequal-variance ANOVA**.

Adjusting for “nuisance variables” when comparing numbers

Sometimes, the variable you’re comparing is influenced not only by which group the subject belongs to, but also by one or more other variables. These variables may not be evenly distributed across the groups you’re comparing (even in a randomized trial).

You can mathematically compensate for the effects of these “nuisance” variables (**confounders**, see **Chapter 20**), by using an **analysis of covariance (ANCOVA)**.

An ANCOVA is like an ANOVA in that it compares the mean value of an outcome variable between two or more groups. But an ANCOVA also lets you specify one or more covariates that you think may influence the outcome.

Comparing Paired Numbers

All the previous tests deal with comparisons between two or more groups of **independent** samples of data. There may be circumstances where you want to compare sets of **paired data**.

Matched-pair data comes up in different flavours as well:

- » The values come from the same subject, but at two or more different times (e.g., before and after some treatment, intervention, or event).
- » The values come from a crossover clinical trial (see **Chapter 28**).
- » The values come from two or more different individuals who have been paired, or matched, in some way (e.g., they may be twins or they may be matched on the basis of having similar characteristics such as age or gender).

Comparing Matched Pairs

Paired comparisons are handled by the **paired student t-test**.

- » If the data are not normally distributed, you can use the nonparametric **Wilcoxon Signed-Ranks test** instead.

Comparing Three or More Matched Numbers

When you have three or more matched numbers, you can use **repeated-measures analysis of variance (RM-ANOVA)**.

- » If the data are not normally distributed, you can use the **nonparametric Friedman test** instead.

Comparing Within-group Changes Between Groups

This is a special situation that comes up very frequently in analyzing data from clinical trials.

One way to analyze this data would be by comparing the changes between the groups with a **one-way ANOVA** (or unpaired t-test if there are only two groups).

Although this approach is statistically valid, clinical trial data are not usually analyzed this way. Instead, almost every clinical trial uses an **ANCOVA** to compare changes between groups.

The Pearson Chi-Square Test

This is the most common statistical test to evaluate **association between two categorical variables**. It's called the chi-square test because it involves calculating a number (a test statistic) that fluctuates in accordance with the **chi-square distribution**.

Based on a 2x2 contingency table, the **H₀ for the chi-square test** asserts that there's no association between the row variable and the column variable.

Now, there are some shortcomings about this statistical test:

» **It's not an exact test.**

- The p value it produces is only approximate, so using $p < 0.05$ as a criterion for significance doesn't necessarily guarantee a 5% Type I error rate. It's accurate when all the cells in the table have large counts, but it becomes unreliable when one or more cell counts is very small (or zero). The simplest rule is that there should be at least five observations in each cell of the table.

» **Isn't good at detecting small but steady progressive trends** across the successive categories of an ordinal variable.

- It may give a significant result if the trend is strong enough, but it's not designed specifically to work with ordinal categorical data.

The Fisher Exact Test

This statistical test gives the exact p-value for tables with large or small cell counts (even cell counts of zero).

The big advantages of the Fisher Exact test are:

- » It gives an exact p-value.
- » It is exact for all tables, with large or small (or even zero) cell counts.

However, there are some shortcomings about this statistical test too:

- » The calculations are a lot more complicated, especially for tables larger than 2x2.
- » The calculations can become numerically unstable for large cell counts, even in a 2x2 table.
- » The exact calculations can become impossibly time consuming for larger tables and larger cell counts.
- » The Fisher Exact test is no better than the chi-square test at detecting gradual trends across ordinal categories.

The Kendall Test

Neither the chi-square nor the Fisher Exact test is designed for testing the association between two ordinal categorical variables. Fortunately, other tests are designed specifically to spot trends in ordinal data.

One of the most common ones involves calculating a test statistic called **Kendall's tau**. The basic idea is to consider each possible pair of subjects, determining whether those two subjects are **concordant** or **discordant** with the **hypothesis that the two variables are positively correlated**.

» For example: if one subject in a pair receives a placebo and was unchanged while the other subject receives a low-dose drug and gets better, that pair would be concordant. But if one subject receives a low-dose drug and got better while another subject receives a high-dose drug and remains unchanged, that pair would be considered discordant.

The **Kendall test** counts how many pairs are concordant, discordant, or noninformative (where both subjects are in the same category for one or both variables). The test statistic is based on the difference between the number of concordant and discordant pairs divided by a theoretical estimate of the standard error of that difference. The test statistic is then looked up in a table of the normal distribution to obtain a **p-value**.

Mantel-Haenszel Chi-Square Test

This is a statistical test to use when analyzing the relationship between two **dichotomous categorical variables**, and you want to control for one or more **confounders**.

Like the chi-square test, the Mantel-Haenszel test is only **an approximation**, and it's most commonly used for 2x2 tables provided the categorical variables are ordinal.

Correlation & Regression

These terms describe a set of statistical techniques to deal with the relationships among variables. They are further detailed in **Chapter 12**.

Key Terms

Define the following terms.

Analysis of variance

Chi-squared test

Fisher's exact test

Friedman test

Kruskal-Wallis test

Mann-Whitney U test

McNemar test

Nonparametric test

Paired data

Parametric test

Pearson correlation test

Spearman correlation test

Statistical test

Student's t-test

Unpaired data

Wilcoxon's rank sum

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Get together with some classmates and identify the type of statistical test more appropriate to be done in the following scenarios:

- Bogton Council decide to see whether performance-related pay would improve morale amongst their lavatory cleaners. Each month, twenty lavatory cleaners are paid on the basis of the length of the bristles on their lavatory brush (on the assumption that the harder they have worked, the shorter their bristles will be). Another twenty are paid their usual near-subsistence-level wages, regardless of how hard they work. After 6 months, each worker is asked to rate how happy they are in their job, using a seven-point scale. Which test would you use to see if performance-related pay has affected workers' morale?
- An experimenter wants to know whether experience affects how well shop-keepers can identify children who ask for cigarettes but are under the legal age for purchasing them. Each of 30 tobacconists is shown a random sequence of 40 photographs of young faces, and asked to decide whether each face is younger or older than the legal

age for buying cigarettes. (Half of the faces are aged above the legal age, and half below). The experimenter records the number of correct decisions per participant, and also asks each shop-keeper how long they have been selling cigarettes. (These latter data turn out to be heavily skewed). Which test should the experimenter use to decide whether experience leads to better age-estimation in this group?

- It's often said that you're hungry again soon after a Chinese meal. An experimenter puts this to the test. There are four conditions, and each participant does each one, on a different day of the week (order of conditions is counterbalanced across participants). In the first, participants eat an Indian takeaway; in the second, they eat a pizza; in the third, they eat a Chinese takeaway; and in the fourth, they eat a Kentucky Fried Chicken takeaway. All the meals are equated for bulk of contents and calorific value. The dependent variable is the loudness of each participants' stomach rumblings (in decibels), measured one hour after they have eaten the meal. These measurements are normally distributed, but much more variable for the "KFC" condition

than the others. Which test should be used to decide whether there is a difference between these meals in terms of how quickly people get hungry again after eating them?

- Some TV viewers complain to the BBC that Jeremy Clarkson's programme "Top Gear" is a bad influence on young drivers, given that it extols the virtues of laddishness, speeding and high performance cars. To determine whether there is any foundation to these claims, a researcher uses a speed camera to measure the speeds of 400 drivers on an A-road, the morning before the programme is transmitted. He follows this procedure again, the morning afterwards. Each car is photographed, so that the experimenter can select only those drivers who traveled that route on both occasions, and hence whose speeds were measured twice. The experimenter subtracts each driver's first speed reading from their second, to get a "difference score": a positive score means a driver drove faster on the second occasion, and a negative score means they drove more slowly. The selected drivers were then contacted and asked whether or not they had watched "Top Gear" that week. Which test would you use to see whether drivers who watched "Top Gear" drove faster the following morning than drivers who did not watch it?
- A researcher is interested in factors affecting reproductive success in *Homo canarywharfensis*, an obscure species of proto-human that inhabits high-altitude habitats in a region of south-east London. Once she has acclimatised them to her presence, she traps a hundred of the males and records the price of their suits. She then releases them back into the wild and follows them for a fortnight, recording how many females each one mates with. Is there a relationship between wealth (as reflected in suit price) and reproductive success (as reflected in how many females each male mates with?) The data for reproductive success are heavily skewed, since most of the males attract no females.

- The local Sussex ale, Harvey's Best bitter, is reputed to be imbued with truly magical medicinal properties, as well as having an especially delicious flavour, a unique golden colour and a beautiful yeasty head. To investigate its effects, a researcher asks four groups of cyclists to cycle up Ditchling Beacon (the highest point on the South Downs). One group drink no Harvey's beforehand; another group drink one pint of Best each; a third group drink two pints each; and a fourth group drink four pints each. The dependent variable is how fast each cyclist gets from the bottom of the Beacon to the top. Which test would you use to see if drinking Harvey's affects the cyclists' speed of ascent?
- It is said that every time someone prints off an email, a penguin dies. To put this to the test, a researcher flies to the South Pole and repeatedly counts the number of penguins, as her colleague at Sussex prints out his emails one at a time. Which test would you use to see if there is a relationship between printing off emails and penguin mortality?
- A researcher investigates four different methods for coping with extreme stress. Each person attempts to assemble an IKEA flat-pack wardrobe (the stress-induction phase of the study), and is then allocated randomly to one of four groups. Those in the first group practice yoga for twenty minutes; those in the second group engage in deep breathing for a similar amount of time; those in the third group spend twenty minutes in a Harvey's pub, drinking Best bitter; and those in the fourth group simply scream at the top of their voice for twenty minutes. Each participant then provides a rating on a 0-10 scale of how stressed they feel. Which test would you use to determine whether the four methods differ in their effectiveness for relieving stress?

- 200 men and 150 women are asked to decide which one of the following features is most important to them when they choose a new car: price, performance, safety level, roominess, or colour. Which test would you use to see if men and women differ in their preferences?
- While sales of traditional classical music CD's are falling, "cross-over" classical performers who sacrifice their integrity for money by producing populist versions of tunes like "Nessun Dorma" are big business. The CD sales of twenty opera singers are examined: ten of these singers are rated as "ugly" by a panel of independent judges, and twenty are rated as "highly attractive". Is the success of these performers related to their physical attractiveness?

Bibliography and Suggested Reading

- Armitage P, Berry G. Statistical Methods in Medical Research. 3rd Edition. Oxford: Blackwell Scientific Publications, 1994.
- Beukelman T, Brunner HI. Trial Design, Measurement, and Analysis of Clinical Investigations. In: Petty RE, Laxer RM, Lindsley CB, Wedderburn LR. Textbook of Pediatric Rheumatology, 7th Edition. Philadelphia: ELSEVIER, 2016.
- Dtsch Arztebl Int. 2010; 107(19): 343–8 DOI: 10.3238/arztebl.2010.0343.
- Kestin I. Statistics in medicine. Anaesthesia and Intensive Care Medicine. 2012;13(4):200–207.
- Nayak BK, Hazra A. How to choose the right statistical test? Indian J Ophthalmol. 2011; 59(2): 85–86.

Correlation and Regression

Learning objectives for this chapter

- A.** Identify the direction and strength of a correlation between two factors.
- B.** Understand the Pearson correlation as a descriptive statistic that measures and describes the relationship between two variables.
- C.** Understand the Spearman correlation and how it differs from the Pearson correlation in terms of data that it uses and the type of relationship that it measures.
- D.** Differentiate the concepts of correlation and causation.
- E.** Interpret the meaning of the correlation coefficient in context.
- F.** Understand the line of best fit as a tool for summarizing a linear relationship and predicting future observed values.
- G.** Understand why the regression line is called the “line of best fit” or “least squares regression”.
- H.** Understanding the many kinds of regression analysis.

Using graphs is a good way to detect relationships between two variables, but sometimes it is not possible to determine this relationship by only looking at a graph. In this scenarios, Statistics play a determinant role.

The words **correlation** and **regression** are often used interchangeably, but they refer to two different things:

- » **Correlation** refers to the **strength** of the relationship between two or more variables.
- » **Regression** refers to a set of techniques for **describing the relationship** between two or more variables.

Correlation

This term was first used by Francis Galton in 1888 in a paper describing the extent to which physical characteristics could be inherited from generation to generation. He said:

“Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction.”

In no more than ten years after this statement, Karl Pearson (the guy who invented the chi-square test) developed a formula for calculating the correlation coefficient from paired values of two variables (X and Y).

The **Pearson correlation coefficient** (represented by the symbol r) measures the extent to which two variables (X and Y) measured on interval or ratio scales, when graphed, **tend to lie along a straight line**.

- » If the variables have **no relationship** (if the points scatter all over the graph), r will be 0.
- » If the relationship is **perfect** (if the points lie exactly along a straight line), r will be $+1$ or -1 .

Correlation coefficients can be **positive** (indicating upward-sloping data) or **negative** (indicating downward-sloping data). **Figure 12.1** shows what several different values of r look like.

How to Analyze a Correlation Coefficient

There are numerous statistical analyses that can be performed on correlation coefficients. We suggest the following:

1. Determine if **r is different from zero**: This can be done by calculating a p -value from the r value.

Keep in mind:

The Pearson correlation coefficient measures the extent to which the points lie along a straight line. So, if your data lies closely along a curved line, the r value may be quite low, or even zero.

Key thing:

In the clinical scenarios, the Pearson correlation coefficient is not as useful as you may think, because the patients (as well as the physicians) are more comfortable to know the **chances** of something happening rather than knowing that **“the data correlate with a number”**.

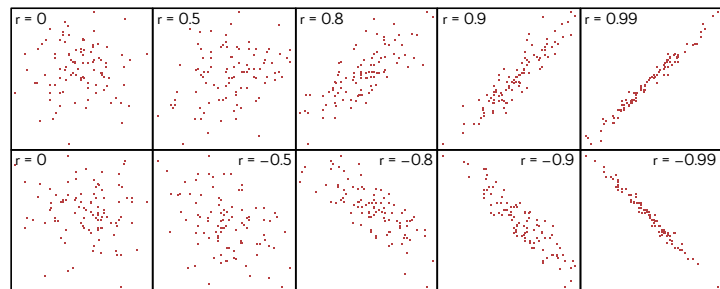


Figure 12.1. Graphic example of 100 data points, with varying degrees of correlation.

2. Calculate confidence limits around an observed r: The z score calculated by the Fisher z transformation can be used to get the lower and upper limits.

- The z score is called a **univariate statistic** because it uses data from a single variable, and conclusions drawn from the statistic must be limited to that variable.
- Suppose that the 95% confidence interval goes from -0.104 to $+0.835$, a range that includes the value **zero**. This means that the true r value could indeed be zero, which would be consistent with a **nonsignificant p-value** (like 0.098) that you obtained from the significance test of r.

3. Determine differences between two r values: By obtaining the z score from each r and then, its standard deviation.

- The test statistic is obtained by dividing the difference by its standard error and can be converted into a p-value.
- If the p-value of is less than 0.05, that means that the two correlation coefficients are **significantly different from each other**.

Types of Correlation

Correlation comes in different flavours:

- » If one or both of the variables are measured on an **ordinal scale**, you **cannot use a Pearson's correlation**.
- » **Spearman's correlation** is very similar to Pearson's correlation, and is intended to analyze **ordinal data**.

How Much is Too Much Correlation?

It is an **arbitrary determination** and must be stated **based on the clinical scenario** dealing with the data.

Correlation and Causation

It's commonly said that "correlation does not equal causation." Although the phrase is true, it is misleading.

The popularity of this phrase incorrectly implies that other statistics do allow you determine causal relationships. In fact, **no statistic, by itself, allows researchers to infer causality**. So, while it is true that correlation does not equal causation, it is also true that the "t-test does not equal causation" and "ANOVA does not equal causation."

To support a causal claim, there are **two requirements**:

- » First, you must establish that **the two variables are significantly related to each other**. In other words, when one variable changes, the other variable changes as well.

Key point:

No statistical relationship equals causation.

» The second requirement is that **there are no confounds or alternative explanations for the statistical association** (more on this on **Chapter 20. Confounding**).

For now, we should understand that no statistic allows us to infer a causal relationship between an indirect variable and a direct variable **unless confounds are controlled**.

Hypothesis Testing and Correlation

Researchers always assume that a null hypothesis is true unless they find sufficient evidence to reject it. In the case of correlation, the null hypothesis asserts that the two variables being studied **are not associated**.

- » If the HO were true, the calculated r value would be **close to 0**.
- » If the calculated r value is far from 0, the null is **not likely to be true**.

Regression

Regression analysis goes beyond just asking whether two (or more) variables are associated; it's concerned with finding out exactly **how** those variables are associated (what formula relates the variables together).

Fitting a formula to a set of data can be **useful in a lot of ways**:

- » You can test for a significant association or relationship between two or more variables (the main reason many researchers do regressions).
- » You can get a compact representation of your data.
- » You can make precise predictions, or prognoses.
- » You can do mathematical manipulations easily and accurately on a fitted function that may be difficult or inaccurate to do graphically on the raw data, that is: interpolate between two measured values or extrapolate beyond the measured range.
- » You can test a theoretical model, such as a multi-compartment kinetic model of a drug's absorption, distribution, metabolism, and elimination from the body.
- » You can obtain numerical values for the parameters that appear in the model.

A **regression model** is usually a formula that describes **how one variable** (called the dependent variable, the response variable, the outcome, or the result) **depends on one or more other variables** (called independent variables, explanatory variables or predictors).

Regression models that have only **one explanatory variable** are known to be a **simple regression model**. If the model has more than one outcomes, it is known to be a **multiple regression model**.

For simple regressions (one predictor and one outcome variable), you can think of the parameters as **specifying the position, orientation, and shape of the fitted line on the scatter plot** (like the slope and Y-intercept of a straight line).

If you have only one independent variable, it is often designated by **X**, and the dependent variable is designated by **Y**. If you have more than one independent variable, variables are usually designated by letters toward the end of the alphabet (W, X, Y, Z).

The general formula of regression analysis is:

$$\text{Outcome} = (\text{model}) + \text{error}.$$

That means that we can predict the outcome based on the model of our data and by adding the error. A large number of models discussed in Medicine are **linear models**, and imply that a straight line through all our data points would fit very well to our data. Further than that, we could have multiple regression lines and try to identify which of those lines fit better to our data. This can be done using the **method of least squares**.

Let's try to better understand regression with the following example.

Suppose we have the data shown in **Figure 12.2-A**. As you can observe, the points (individual data) are spread throughout the graphic. If we obtain the mean of the data, we could calculate the deviations of the data points with respect of the mean. These deviances give information about how well the mean fits in the data. In regression, these deviances are called **residuals**.

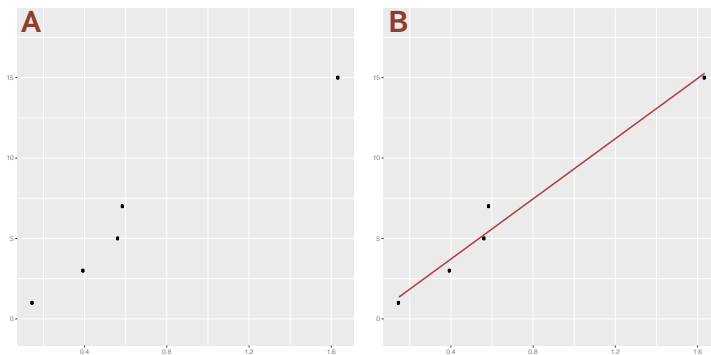


Figure 12.1. Graphic example of 100 data points, with varying degrees of correlation.

Now let's draw a possible regression line in our data and calculate the residuals (**Figure 12.2–B**).

Regression allows us to select **the line with the lowest residuals** (the lowest sum of squared differences).

In order to assess how good the regression line fits the data, we sum the differences between the mean and the data points (**total sum of squares [SST]**). While the least squares method helped us to find the best regression line for our data, this model has some amount of **error**. If we square this error (differences) and sum them, we obtain a value known as the **residual sum of squares (SSr)**. If we rest the SSr from the SST, we obtain the **sum of squares of our model (SSM)**.

Now, **please don't get lost in here!** That last paragraph is only to justify the following one. You **don't need to memorize everything!!**

If we divide SSM by SST, we obtain a value called **R²** that can be expressed as a percentage if multiplied by 100. This value **represents the percentage of the variation in the outcome variable that can be explained by the model**.

Why Using Regression?

Regression is used because, in real life, there is a lot of **biological diversity**, and the data obtained from that diversity will never ever fit in a single-drawn line.

Types of Regression

Broadly speaking, you can classify regression on the basis of:

- » **How many outcomes** (dependent variables) appear in the model:
 - Univariate regression has only one outcome variable.
 - Multivariate regression has ≥ 2 outcome variables.
- » **How many predictors** (independent variables) appear in the model:
 - Univariable regression has only one predictor variable.
 - Multivariable regression has ≥ 2 predictor variables.
- » What **kind of data** the outcome variable is:
 - **Ordinally regression** is used when the outcome is a continuous variable whose random fluctuations are governed by the normal distribution.
 - **Logistic regression** is used when the outcome variable is a dichotomous category whose random fluctuations are governed by the binomial distribution.

- **Poisson regression** is used when the outcome variable is the number of occurrences of a sporadic event whose random fluctuations are governed by the Poisson distribution.
- **Survival regression** is used when the outcome is a time to event (survival time).
- » What kind of **mathematical form** the model takes:
 - In a **linear function** you multiply each predictor variable by a parameter and then add these products to give the predicted value. When you gave one parameter that isn't multiplied by anything, it's called the constant term or the intercept.
 - In a **nonlinear function** you find anything that's not a linear function.

Key Terms

Define the following terms.

Causation	R²	Spearman's correlation
Correlation	Regression	Types of regression
Method of least squares	Regression model	
Pearson correlation	Residuals	

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- 1. Answer:** When is it appropriate to use a Pearson's correlation analysis?
- 2. Answer:** How do you tell the direction of the relationship?
- 3. Answer:** Under what circumstances should you use a Spearman's rho correlation analysis rather than a Pearson's?
- 4. Match the following correlation coefficients with the scatterplots shown in Figure AL12.1.**
 - $r = 0.73$:
 - $r = 0.87$:
 - $r = -0.42$:
 - $r = -0.77$:

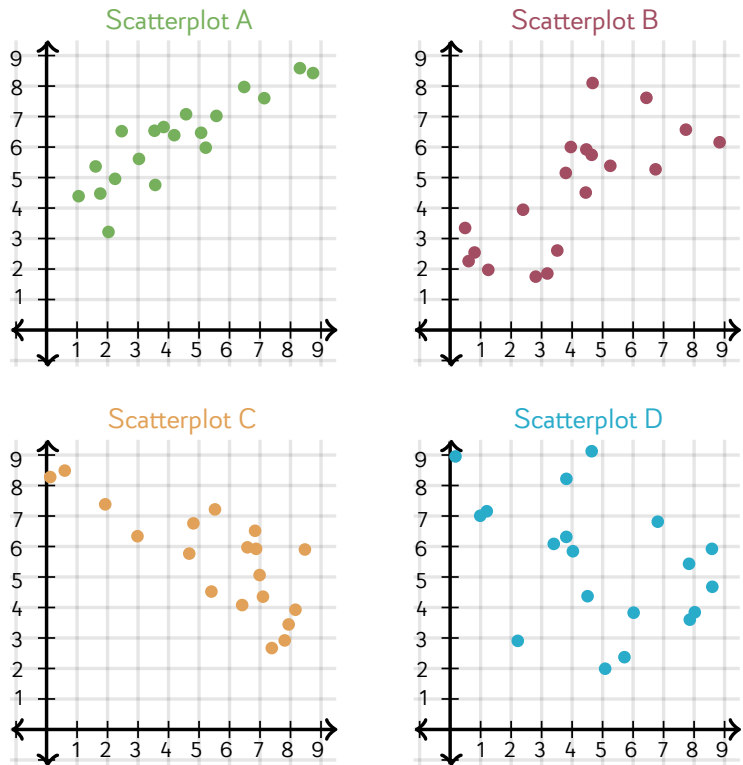


Figure AL12.1. Scatterplots for exercise 4.

5. Multiple choice questions.

1. The correlation coefficient is used to determine:

- a) A specific value of the y-variable given a specific value of the x-variable.
- b) A specific value of the x-variable given a specific value of the y-variable.
- c) The strength of the relationship between the x and y variables.
- d) All of the above.
- e) None of the above.

2. What type of data is required for a Pearson's analysis which does not include a dichotomous variable?

- a) Ratio or nominal.
- b) Categorical or ratio.
- c) Nominal or ordinal.
- d) Interval or nominal.
- e) Interval or ratio.

3. If all points cluster in an ascending line this would suggest what?

- a) There would be no significant relationship.
- b) There would be a weak positive relationship.
- c) There would be a strong negative relationship.
- d) There would be a strong positive relationship.
- e) There would be a non-linear relationship.

4. If most points depict a dispersed descending line this would suggest what?

- a) There would be no significant relationship.
- b) There would be a weak positive relationship.
- c) There would be a strong negative relationship.
- d) There would be a strong positive relationship.
- e) There would be a weak negative relationship.

5. A Pearson test statistic of 0.876 with a significance level of $P < 0.01$ would suggest what?

- a) This would suggest that there is a significant, strong, positive relationship.
- b) This would suggest that there is a significant, weak, positive relationship.
- c) This would suggest that there is a non-significant, weak, negative relationship.
- d) This would suggest that there is a non-significant, weak, positive relationship.
- e) This would suggest that there is a significant, strong, negative relationship.

6. When reporting Pearson's correlation coefficient, what letter do you use to indicate which test you used?

- a) x .
- b) Σ .
- c) r .
- d) P .
- e) N .

7. What correlation can you use if your data do not meet the assumptions of normal distribution?

- a) Mixed ANOVA.
- b) Spearman's rho.
- c) Chi-square.
- d) Paired t-test.
- e) Independent samples t-test.

8. A Spearman's test statistic of -0.207 with a $p = 0.057$ would suggest what?

- a) This would suggest a strong, negative relationship which is approaching significance.
- b) This would suggest a strong, significant, positive relationship.
- c) This would suggest a weak negative relationship which is approaching significance.
- d) This would suggest a weak, non-significant, positive relationship.
- e) This would suggest a weak, non-significant, negative relationship.

9. When reporting a Spearman's correlation, what letter do you use to indicate which test you used?

- a) x .
- b) Σ .
- c) r .
- d) P .
- e) N .

10. What does a partial correlation analysis allow you to do which Pearson's and/or Spearman analyses do not?

- a) It allows you to use interval data.
- b) It allows you to use data which is not normally distributed.
- c) It allows you to control covariates.
- d) It allows you to use ratio data.
- e) It allows you to use dichotomous variables.

11. If there is a very strong correlation between two variables then the correlation coefficient must be:

- a) Any value larger than 1.
- b) Much smaller than 0 if the correlation is negative.
- c) Much larger than 0 regardless of whether the correlation is negative or positive.
- d) Any value smaller than 1.
- e) None of the alternatives is correct

12. In regression, the equation that describes how the response variable (y) is related to the explanatory variable (x) is:

- a) The correlation model.
- b) The regression model.
- c) Used to compute the correlation coefficient.
- d) The Pearson's method.
- e) The Spearman's method.

13. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study:

$$y = -0.0127 + 0.0180x$$

The above equation implies that:

- a) A each beer consumed increases blood alcohol by 1.27%.
- b) On average it takes 1.8 beers to increase blood alcohol content by 1%.
- c) Each beer consumed increases blood alcohol by an average of amount of 1.8%.
- d) Each beer consumed increases blood alcohol by exactly 0.018.

14. Regression modeling is a statistical framework for developing a mathematical equation that describes how:

- a) One explanatory and one or more response variables are related.
- b) One response and one or more explanatory variables are related.
- c) Several explanatory and several response variables response are related.
- d) All of these are correct.

15. Regression analysis was applied to return rates of immigrants to their country. Regression analysis was used to study the relationship between return rate (x : % of people that return to the country in a given year) and immigration rate (y : % of new adults that join the country per year). The following regression equation was obtained:

$$y = 31.9 - 0.34x$$

Based on the above estimated regression equation, if the return rate were to decrease by 10% the rate of immigration to the country would:

- a) Increase by 34%.
- b) Increase by 3.4%.
- c) Decrease by 0.34%.
- d) Decrease by 3.4%

16. In regression analysis, the variable that is used to explain the change in the outcome of an experiment, or some natural process, is called:

- a) The x -variable.
- b) The independent variable.
- c) The predictor variable.
- d) The explanatory variable.
- e) All of the above.

17. A correlation coefficient tells us:

- a) The slope of the equation through the data.
- b) How well one variable predicts another.
- c) How closely two variables are linearly related.
- d) Whether the relationship is linear or non-linear.

18. A correlation coefficient is a reliable estimate of association if:

- a) A large sample size has been enrolled.
- b) The sample is randomly selected from the population.
- c) Extra cases are enrolled to ensure a wide range of y values.
- d) There is a positive association between two variables.

19. A regression model is more reliable if:

- a) It has been created using a statistical package.
- b) There is only one explanatory variable.
- c) One of the explanatory variables is a binary characteristic.
- d) The explanatory variables are not related to one another.

20. If two variables, x and y, have a very strong linear relationship, then:

- a) There is evidence that x causes a change in y.
- b) There is evidence that y causes a change in x.
- c) There might not be any causal relationship between x and y.
- d) None of these alternatives is correct.

Bibliography and Suggested Reading

- Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. *Crit Care*. 2003 Dec;7(6):451-9.
- Crawford SL. Correlation and Regression. *Circulation*. 2006;114:2083–2088.
- Esparza-Villalpando V. *Basic Biostatistics with R*. MCIC. San Luis Potosi; 2019.
- Pezzullo JC. *Biostatistics for Dummies*. Hoboken, NJ: John Wiley & Sons, Inc; 2013.
- Poldrack RA. *Statistical Thinking for the 21st Century*. Available at: <http://statstinking21.org>.
- Scheff SW. Correlation and Regression. In: Scheff SW. *Fundamental Statistical Principles for the Neurobiologist: A Survival Guide*. London: ELSEVIER; 2016.

Section IV

Risk and Prognosis of Diseases

Chapters of the Section

Chapter 13	Study Designs
Chapter 14	Cohort Studies
Chapter 15	Case-control Studies
Chapter 16	Cross-sectional Studies
Chapter 17	Survival Analysis
Chapter 18	Disease Occurrence, Risk, Association, Importance, and Implication
Chapter 19	Odds Ratio and Relative Risk: As Simple as It Can Get
Chapter 20	Confounding
Chapter 21	Attributable Risk

Study Designs

Learning objectives for this chapter

- A. Define a study design.
- B. Understand and describe the four main axes of the architecture of a study design.
- C. Identify a study design by looking at three key issues.

One of the most important considerations in Epidemiology is the study design. The way a study is **framed** determines what kind of information will be collected, how that information will be used, and what types of measurements will be calculated.

Epidemiological studies can be divided into different groups (**Figure 13.1** and **Figure 13.2**), but three types predominate:

- » **Descriptive.**
- » **Analytic** (observational).
- » **Experimental** (interventional).

Classification Criteria

The most important characteristics of the architecture of a study can be classified according to the following **four main axes** (**Table 13.1**):

- » **Purpose of the study:** Analytical or descriptive.
- » **Temporal sequence:** Transverse or longitudinal.
- » **Control of the assignment of the study factors:** Experimental or observational.
- » **Start of the study in relation to the chronology of events:** Prospective or retrospective.

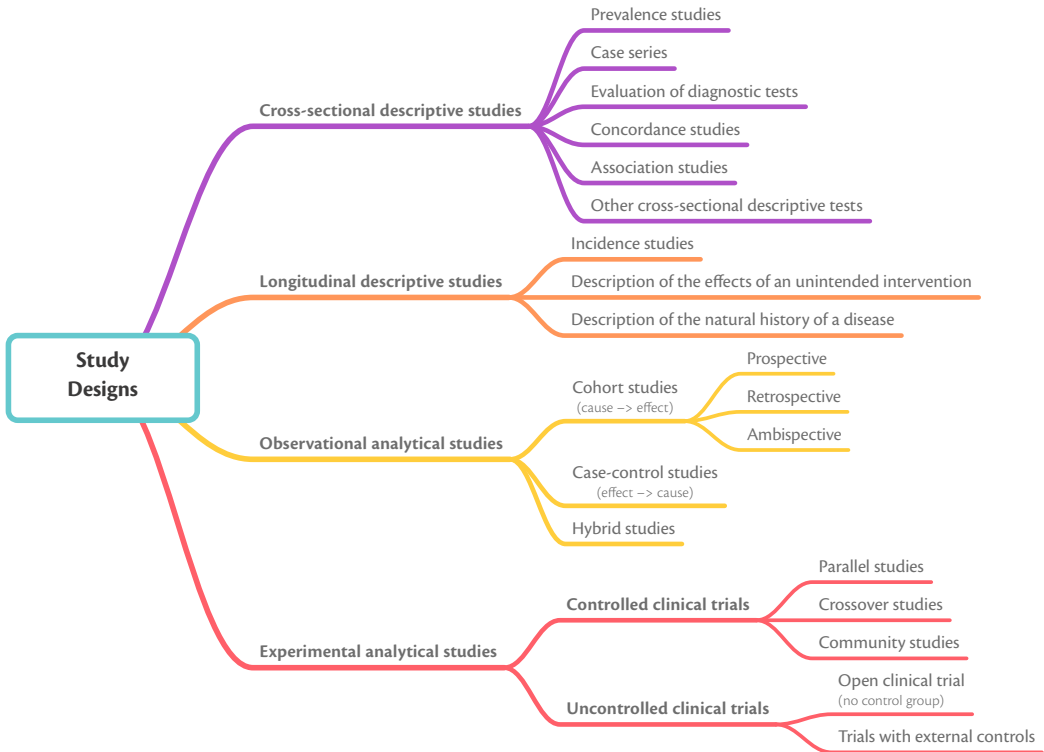


Figure 13.1. Types of Study Designs.

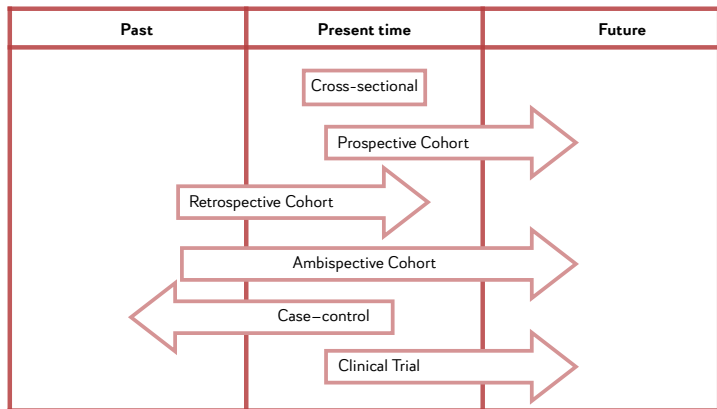


Figure 13.2. Temporal direction of Study Designs. The arrow implies longitudinality, and its direction represents the temporal direction of the study.

Table 13.1. Main axes of the architecture of a study

Axe	Characteristics
Purpose	<p>A study is considered analytical when its purpose is to evaluate an alleged causal relationship between a factor and an effect, an answer or a result</p> <p>A study is considered descriptive when it does not seek to evaluate an alleged cause-effect relationship, but its data are used for purely descriptive purposes</p> <p>Descriptive studies are useful to generate etiological hypotheses that should be contrasted later with analytical studies</p>
Temporal sequence	<p>A study is considered transverse (cross-sectional) when the data of each subject essentially represent a moment of time. These data may correspond to the presence, absence or different degrees of a characteristic or a disease or may examine the relationship between different variables in a defined population at a given time</p> <ul style="list-style-type: none"> • The variables are measured simultaneously, therefore we cannot establish a time sequence between them and we cannot infer a cause-effect relationship • This studies are by definition, descriptive <p>A study is considered longitudinal when there is a time span between the different variables that are evaluated, allowing us to establish a time sequence between them</p> <ul style="list-style-type: none"> • Longitudinal studies can be both descriptive and analytical • In the analytical longitudinal studies, the temporal direction must be taken into account, which can go from the cause to the outcome (experimental studies and cohort studies) or from the outcome to the cause (case and control studies) • A study is considered to be longitudinal if the observations refer to two moments in time, even when the collection of information has been carried out simultaneously
Assignment of the study factors	<p>A study is considered experimental when the research team assigns the study factor and deliberately controls it for conducting the investigation according to a pre-established plan</p> <ul style="list-style-type: none"> • These studies focus on a cause-effect (analytical) relationship, and generally assess the effect of one or more preventive or therapeutic interventions <p>A study is considered observational when the study factor is not controlled by the researchers, and these are limited to observe, measure and analyze certain variables in the subjects</p>
Start of the study in relation to the chronology of the events	<p>A study is considered prospective if its beginning is prior to the studied facts, so that the data are collected as they happen</p> <p>A study is considered retrospective if its design is subsequent to the studied facts, so that the data are obtained from files or records</p> <p>A study is considered ambispective if it has a combination of both situations</p>

Spotting the Study Design

The type of study can be identified by looking at **three issues** (as per the Tree of designs in **Figure 13.1**):

1. What was the aim of the study?

- » To simply describe a population.
 - Descriptive.
- » To quantify the relationship between factors.
 - Analytic.

2. If analytic, was the intervention randomly allocated?

- » Yes?
 - Randomized Clinical Trial (RCT).
- » No?
 - Observational study.
 - For observational studies, the main types will then depend on the timing of the measurement of outcome. This brings us to the third question:

3. When were the outcomes determined?

- » Some time after the exposure or intervention?
 - Cohort study (prospective study).
- » At the same time as the exposure or intervention?
 - Cross sectional study.
- » Before the exposure was determined?
 - Case-control study (retrospective study).

Key Terms

Define the following terms.

Ambispective study

Analytical study

Cross-sectional study

Descriptive study

Experimental study

Longitudinal study

Observational study

Prospective study

Retrospective study

Study design

Transverse study

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Read the descriptions of the following studies and determine the study design used.

- Cryptosporidiosis is an enteric illness that is frequently waterborne. A research team found no published studies of the risk factors for cryptosporidiosis in immunocompetent adults. They recruited patients with cryptosporidiosis from a surveillance system, and age-matched controls were recruited by random-digit dialing. Subjects in both groups were interviewed by telephone to obtain information about previous exposures.
- A study to determine the efficacy of immunotherapy with ant venom for treating ant stings involved a group of 68 adults who were allergic to ant stings; each subject was randomly assigned to receive either venom immunotherapy or a placebo. After a sting challenge in which any reactions were recorded, the group originally on the placebo was given the venom immunotherapy, and after a sufficient time, they too were given a sting challenge.
- A Cancer Outcomes Study was designed to investigate the patterns of prostate cancer care and effects of treatment on quality of life. Eligible cases were identified from pathology facilities within 6 months of diagnosis. A random sample of eligible cases were contacted and asked to complete a questionnaire on their initial treatment and to provide permission to the investigators to abstract their medical records to obtain information on their initial care.
- A study aiming to investigate factors contributing to medical students' self-perceived competency in cancer screening examinations. Students were asked to assess their competency in performing several cancer screening examinations, and multiple regression analysis was used to identify predictors of competency.
- A study to determine whether treatment with a calcium channel block or an angiotensin-converting enzyme inhibitor lowers the incidence of coronary heart disease when compared with a diuretic included over 33,000 patients. The primary outcome was fatal coronary heart disease or myocardial infarction.
- Questionnaires were mailed to every 10th person listed in the city telephone directory. Each person was asked to list age, sex, smoking habits, and respiratory symptoms during the preceding seven days. About 20% of the questionnaires were completed and returned. About 10% of respondents reported having upper respiratory symptoms.
- 1,500 employees of a major aircraft company were initially examined in 1951 and were classified by diagnostic criteria for coronary artery disease (CAD). New cases of CAD have been identified by examinations every three years and through death certificates. Attack rates in different subgroups have been computed.
- A random sample of middle-aged sedentary adults were selected from four census tracts, and each person was examined for coronary artery disease (CAD). All persons without disease were randomly assigned to either a two-year program of aerobic exercise or a two-year arthritis-prevention non-aerobic exercise program. Both groups were observed semi-annually for incidence of CAD.
- A 39-year old woman presents with a mild sore throat, fever, malaise and headache and is treated with penicillin, for presumed streptococcal infection. She returns in a week with hypertension, fever, rash and abdominal pain. She responds favorably to chloramphenicol, after a diagnosis of Rocky Mountain spotted fever is made.

2. The abuse of Phenacetin, a common ingredient of analgesic drugs, can lead to kidney disease. There is also evidence that use of salicylate provides protection against cardiovascular disease.

- How would you design a study to examine the effects of these two drugs on mortality due to different causes and on cardiovascular morbidity?

3. Select a study with an interesting topic. Carefully examine the research question and decide which study design would be optimal to answer the question.

- Is that the study design used by the investigators?
- If so, were the investigators attentive to potential problems identified in this chapter?
- If not, what are the reasons for the study design used? Do they make sense?

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4^o edición. Barcelona: ELSEVIER; 2013.
- Centre for Evidence-Based Medicine. Study Designs. 2016. [Last accessed on 2020 Mar 04]. Available from: <https://www.cebm.net/2014/04/study-designs/>.
- Forthofer RN, Lee ES, Hernandez M. Biostatistics. 2^o Edition. Burlington: Elsevier Academic Press; 2007.
- Parfrey PS, Barret BJ. Clinical Epidemiology. Practice and Methods. 2^o Edition. New York: Springer; 2015.
- Ranganathan P, Aggarwal R. Study designs: Part 1 – An overview and classification. *Perspect Clin Res*. 2018 Oct-Dec; 9(4): 184–186.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. *Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication*. Waltham: Elsevier Academic Press; 2015.

Cohort Studies

Learning objectives for this chapter

- A. Determine what a cohort means.
- B. Describe the architecture of the cohort studies.
- C. Describe the types of epidemiological parameters that can be estimated with cohort studies.
- D. State the advantages and disadvantages of cohort studies.
- E. Identify the main classification of cohort studies.
- F. Learn to apply the main features of a cohort study.

The term “cohort” is derived from the Latin word *cohors*. Roman legions were composed of 10 cohorts. During battle, each cohort, or military unit, consisting of a specific number of warriors and commanding centurions, were traceable. The word **cohort** has since been adopted into Epidemiology to define a set of people that have one characteristic or a set of characteristics in common (usually the exposure to a study factor) that will be followed over a period of time.



A **cohort** is group of individuals who share a common trait, that is part of a clinical trial or study, and that is observed over a period of time.

A cohort may correspond to:

- » **A generation:** People defined by the same date of birth.
- » **A professional group:** Doctors from a country.
- » **People who have a certain exposure in common:** Women treated for breast cancer.
- » **People who have a genetic characteristic in common:** Trisomy 21.
- » **A geographically defined community:** The inhabitants of the San Luis Potosí population.

A cohort study is a **longitudinal**, **analytical**, and **observational** design that **compares two cohorts**, or two groups within the same cohort, that **differ in their exposure to the study factor**, with the aim of **assessing a possible cause-effect relationship**.

In a cohort study, individuals without the disease or the effect of interest are arranged into groups based on their exposure or not to the study factor. Those groups are followed over a period of time, comparing the **frequency with which the effect or response appears in those exposed and unexposed**.

Advantages and disadvantages of the Cohort Studies are summarized in **Table 14.1**. Possibly the greatest limitation of the Cohort Studies is that it **cannot establish causality**. In order to determine causation, **Bradford Hill's criteria** must be met (**Appendix B**).

Types of Cohort Studies

Prospective cohort study

These studies are carried out **from the present time into the future** (**Figure 13.2**). The researcher starts from the formation of the groups of subjects exposed and not exposed to a possible risk factor, and follows them for a while.

This type of studies is invaluable exemplified by the landmark **Framingham Heart Study**, started in 1948 and still ongoing.

- » **Strengths:** Powerful tool to assess incidence, helpful in investigating the potential causes of the condition, allows to measure variables more completely than retrospectively.
- » **Weaknesses:** All cohort studies are observational studies, therefore causal inference is challenging and interpretation is often muddled by confounders.

Table 14.1. Advantages and Disadvantages of Cohort Studies

Advantages	Disadvantages
<ul style="list-style-type: none"> • They allow the direct calculation of the incidence rate in the exposed and unexposed cohort • They allow the calculation of the relative risk of those exposed in relation to those not exposed • They ensure an adequate time sequence between the study factor and the outcome • They allow to evaluate the effects of the risk factor on various diseases 	<ul style="list-style-type: none"> • They are not efficient for the study of rare diseases • They are not efficient for the study of diseases with long latency periods • They require a large number of participants • They have high cost (prospective design)

Retrospective cohort study

These studies are carried out **at the present time and look back into the past** (Figure 13.2). Both exposure and disease have already occurred, and the identification of the exposed and unexposed cohorts is based on their situation on a well-defined prior date.

- » **Strengths:** Same as the prospective cohort study, and they have the advantage of being much less costly and time consuming.
- » **Weaknesses:** There is a limited control of the investigator over sampling the population and over the nature and quality of predictor variables.

Ambispective cohort study

In these studies, data are collected **retrospectively and prospectively** in the same cohort (Figure 13.2).

Methodological Pearls in Cohort Studies

- » The **hallmark** of a cohort study is defining the selected group of subjects **by exposure status** at the start of the investigation.
- » Both exposed and unexposed groups must be selected **from the same population**.
- » Because prospective cohort studies may require a long follow-up, **losses must be minimized** (loss to follow-up rate should not exceed 20 % of the sample).

Data that can be Obtained from Cohort Studies

Given the fact that cohort studies are **longitudinal**, we can make the following estimates (Table 14.2):

- » The **incidence** of the disease in the exposed and unexposed groups.
- » The **relative risk** of the association between the risk factor and the outcome variable.
- » The **fraction or proportion attributable** or proportion of cases of a disease that results from exposure to a particular factor or a combination of them.
- » The **difference in incidences** as a measure of the potential impact that the elimination of exposure would have.



Disease Occurrence and Risk

Incidence: The number of new cases of a condition that develop in a population during a defined time period.

Relative risk: Ratio of the probability of the outcome occurring in the exposed group divided by the probability of the outcome occurring in the non-exposed group.

Table 14.2. Summary of Differences Between Case-control Studies, Cross-sectional Studies, and Cohort Studies

Category	Case-control	Cross-sectional	Cohort
Sample	Starts with ill subjects (cases) and healthy subjects (controls)	All subjects are included	Starts with healthy subjects
Measures of occurrence			
Incidence	No	No	Yes
Prevalence	No	Yes	No
Measures of association			
	OR	RR OR	RR OR IRR

Key Terms

Define the following terms.

Ambispective cohort study

Cohort

Difference in incidences

Fraction or proportion attributable

Incidence

Prospective cohort study

Relative risk

Retrospective cohort study

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- 1. Answer:** Which study design is most appropriate for describing the incidence and natural history of a health-related event?
- 2. What is the distinction between prospective and retrospective cohort studies?**
- 3. State the features of a cohort study design.**
- 4. List some important advantages to using the cohort study design.**
- 5. List some important disadvantages to using the cohort study design.**
- 6. List the estimates that can be obtained with a cohort study.**

- 7. Draw a diagram of the cohort study design.**
- 8. Complete Table AL16.1 with the description, strengths, and weaknesses of observational analytic study designs.**
- 9. Multiple choice questions.**
- 1. Which of the following is not a key feature of a cohort study?**
- Retrospectively or prospectively, the investigators identify a cohort of subjects who initially did not have the disease or outcome of interest.
 - The groups being compared differ in their exposure status.
 - Investigators measure and compare the incidence of disease among different exposure groups.
 - It is essential that follow up is complete in all subjects.
- 2. An ambispective study is best described as:**
- A study with inconsistent results.
 - A study in which the subjects have equal dexterity with their right and left hand.
 - A study with some components that are like a case-control study and other components that are like a retrospective cohort study.
 - A study with some components that are like a prospective cohort study and other components similar to a retrospective cohort study.
- 3. Cohort studies are good for studying rare diseases.**
- True.
 - False.
- 4. Loss to follow-up rate should not exceed which percentage of the sample?**
- 10%.
 - 15%.
 - 20%.
 - 25%.
- 5. A cohort study can assess causality.**
- True.
 - False.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Coggon D, Rose G, Barker DJP. Epidemiology for the uninitiated. 4ª Edition. London: BMJ Publishing Group Ltd; 2020. Available at: <https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated>.
- Johnson LL. Design of Observational Studies. In: Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3ª Edition. London: Elsevier Academic Press; 2012.
- Parfrey PS, Barret BJ. Clinical Epidemiology. Practice and Methods. 2ª Edition. New York: Springer; 2015.
- Song JW, Chung KC. Observational Studies: Cohort and Case-Control Studies. *Plast Reconstr Surg*. 2010 Dec; 126(6): 2234–2242.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.

Case-control Studies

Learning objectives for this chapter

- A. Differentiate between a case and a control group.
- B. Describe the architecture of the case-control studies.
- C. Describe the types of epidemiological parameters that can be estimated with case-control studies.
- D. State the advantages and disadvantages of case-control studies.
- E. Identify the main classification of the case-control studies.
- F. Learn to apply the main features of a case-control study.

Case-control studies were historically borne out of interest in determining the cause of disease. We could say that the conceptual basis of the case-control study is **similar to taking a history and physical examination of the patient**: the patient with a certain disease is questioned and examined, and elements from this history taking are knitted together to reveal characteristics or **factors that predisposed the patient to the disease**.

In this type of studies, once the **outcome of interest** is chosen (e.g., the patient undergone a specific type of surgery, experienced a complication, has been diagnosed with a disease, etc), **two groups are chosen**:

- » A group of individuals with a specific disease (**cases**).
- » A group of individuals in which the specific disease is absent (**controls**).

Study patients who have developed a disease are identified and their **past exposure to suspected etiological factors** is compared with that of controls who do not have the disease.

As such, data regarding exposure to a risk factor or several risk factors are collected **retrospectively** (Figure 13.2), typically by interview, from records, or with surveys.

Advantages and disadvantages of the case-control studies are summarized in **Table 15.1**.

Table 15.1. Advantages and Disadvantages of Case-control Studies

Advantages	Disadvantages
<ul style="list-style-type: none"> • Good for examining rare outcomes or outcomes with long latency • Relatively quick to conduct • Relatively inexpensive • Requires comparatively few subjects • Existing records may be used • Multiple exposures or risk factors can be examined 	<ul style="list-style-type: none"> • Susceptible to recall bias or information bias • Difficult to validate information • Control of extraneous variables may be incomplete • Selection of an appropriate comparison group may be difficult • Rates of disease in exposed and unexposed patients cannot be determined

Selection of Cases

In order to identify the patients that will conform the cases group, the **definition of the disease**, as well as the **criteria that must be met** by those who present the disease, must be clearly and explicitly established. On the other hand, the **selection criteria** should be aimed at only including **patients who have potentially been exposed** to the alleged risk factor and should be applied equally to both cases and controls groups.

A **selection bias** appears when cases or controls are included or excluded from a study due to some characteristic related to the exposure.

Selection of Controls

The selection of the control group tends to be more problematic because controls must satisfy **two requirements** –which often it proves impossible to satisfy:

- » Their exposure to risk factors and confounders should be **representative of that in the population “at risk”** of becoming cases - that is, people who do not have the disease under investigation, but who would be included in the study as cases if they had.
- » The exposures of controls should be **measurable with similar accuracy** to those of the cases.

Sources to Select Controls

- » **General population:** Their exposures are likely to be representative of those at risk of becoming cases; however, their exposures may not be comparable with that of cases.

» **Patients with other diseases:** especially if subjects are not told the exact focus of the investigation; however, their exposures may be unrepresentative.

Types of Case-Control Studies: Hybrid Designs

Hybrid designs have characteristics of **both cohort studies and case-control studies**, but obviate some of their disadvantages. They analyze **all cases** that appear **in a stable cohort** followed in time and use as only a **sample** of the subjects of that same cohort as a control group.

Depending on the sampling plan used to establish the groups from the components of the cohort, **two general types of designs can be distinguished:** cohort and case studies and case-control nested within a cohort.

Case-control Nested Within a Cohort

From a cohort study already carried out, or that is being carried out, all the subjects that have developed the disease are identified, which will constitute the **case group**. When a case appears, one or more controls are randomly selected **among the subjects at risk at that time**.

Controls can be matched with cases, and it is convenient to do so by some time-dependent variable, such as the years of stay in the cohort. In addition, the same subject could be selected as a control on more than one occasion for different cases, or it could be selected as a control at a given time and considered as a later case if the disease develops.

This design is used when it is necessary to **make very expensive measurements**.

Cohort and Case Studies

A sample of the initial cohort (called “**subcohort**”) is randomly selected. It will serve as a **comparison group for all cases that appear during the follow-up** of the study, regardless of whether or not they already belong to the subcohort. In other words, all cases of the initial cohort that appeared during the follow-up are chosen, and their information is compared with a sample of the initial cohort. The aim is to obtain a new cohort, with fewer subjects than the initial one, in which **cases are overrepresented**.

The same subcohort can serve as a comparison group for the study of various diseases.

This design allows to determine the **incidence rates of the disease**, and not only the relative risk.

Disease Occurrence and Risk

Odds ratio: Ratio of the odds of the outcome occurring in one group divided by the odds of the outcome occurring in another group.



Data that can be Obtained from the Case-Control Studies

The **estimates** that can be obtained are the **proportion of cases and controls** exposed to a possible risk factor. It's also of interest the intensity and duration of the exposure in each of the groups.

The **measure of association** or the risk of suffering a certain health problem associated with the presence of an exposure is called "**odds ratio**" (**OR**) (**Table 14.2**).

Key Terms

Define the following terms.

Case-control Nested Within a Cohort Cases

Cohort and case studies Controls

Odds ratio Proportion of cases and controls

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. What is the distinction between cases and controls?
2. State the features of a case-control study design.
3. List some important advantages to using the case-control study design.
4. List some important disadvantages to using the case-control study design.
5. List the estimates that can be obtained with a case-control study.
6. Draw a diagram of the case-control study design.
7. Discuss the advantages and disadvantages of hospital, general population, and special population controls in a case-control study.
8. Complete **Table AL16.1** with the description, strengths, and weaknesses of observational analytic study designs.
9. Multiple choice questions.
 1. How does the strategy for a case-control study differ from that of a cohort study?
 - a) Case-control studies are retrospective, while cohort studies are always prospective.
 - b) Randomization can be used in a cohort study, but can't be used in a case-control study.
 - c) In case-control studies subjects are selected and grouped based on their disease status, but in cohort studies subjects are selected and grouped based on exposure status.
 - d) The goal of cohort studies is to test an association, but case-control studies just document the frequency of risk factors.

2. In a case-control study one can calculate either a risk ratio or an odds ratio.

- a) True.
- b) False.

3. What is the main reason why it is important to use precise, specific criteria for what constitutes a “case,” i.e. in defining the outcome?

- a) To avoid misclassification with respect to the outcome.
- b) To limit the number of subjects in the study.
- c) To avoid selection bias.
- d) To avoid interviewer bias.

4. Which of the following is not an advantage to case-control studies?

- a) They tend to be less expensive and more efficient than prospective cohort studies.
- b) They are feasible for rare diseases.
- c) Selection of an appropriate comparison group is easy to achieve.
- d) They are good for diseases that have a long latency period (i.e., a long time between exposure and manifestation of disease.).

5. In a recent matched case-control study, 200 cases with hepatocellular carcinoma were individually matched to 200 controls without hepatocellular carcinoma by sex and age (± 5 years). The investigators collected information, for each subject, on a number of potential risk factors and were interested in determining which of them was associated with hepatocellular carcinoma. Which one of the following statements is true?

- a) The authors should use conditional logistic regression methods to analyze the outcomes from this study.
- b) The study investigators decided to match cases and controls by age and sex as they were particularly interested in the associations between each of these variables and hepatocellular carcinoma.
- c) When calculating the odds ratios, the study investigators should ignore the fact that the cases and controls are matched by age and sex.
- d) Had the authors loosened their matching criteria to ensure that cases and controls were matched by age within 10 rather than 5 years, the results from the study would have been strengthened.
- e) The authors should use multiple linear regression methods to analyze the outcomes from this study.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Coggon D, Rose G, Barker DJP. Epidemiology for the uninitiated. 4ª Edition. London: BMJ Publishing Group Ltd; 2020. Available at: <https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated>.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.
- Johnson LL. Design of Observational Studies. In: Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3ª Edition. London: Elsevier Academic Press; 2012.
- Song JW, Chung KC. Observational Studies: Cohort and Case-Control Studies. *Plast Reconstr Surg*. 2010 Dec; 126(6): 2234–2242.
- Parfrey PS, Barret BJ. Clinical Epidemiology, Practice and Methods. 2ª Edition. New York: Springer; 2015.

Cross-sectional Studies

Learning objectives for this chapter

- A. Describe the architecture of the cross-sectional studies.
- B. Describe the types of epidemiological parameters that can be estimated with cross-sectional studies.
- C. State the advantages and disadvantages of cross-sectional studies.
- D. Identify the main uses of cross-sectional studies.

Cross-sectional studies make observations about the presence of diseases, conditions, or health-related characteristics in a **defined population at a specific point in time**. Hence, there is no time dimension involved, as all data are collected at or around the time of the investigation.

Such information can be used to explore etiology. However, associations must be interpreted with caution. **Bias** may arise because of selection into or out of the study population. A cross-sectional design may also make it difficult to establish what is cause and what is effect.

Unlike case–control studies (where participants are selected based on the outcome status) or cohort studies (where participants are selected based on the exposure status), the participants in a cross-sectional study are just **selected based on the inclusion and exclusion criteria** set for the study. Once the participants have been selected, the investigator follows the study to assess the exposure and the outcomes.

Uses of Cross-sectional Studies

A cross-sectional study may be used:

- » For population-based surveys.
- » For estimating the prevalence in clinic-based studies.
- » To calculate the OR.
- » For planning health care.

Disease Occurrence and Risk



Prevalence: The total number of people in a population with a condition at a given point in time.

Odds ratio: Ratio of the odds of the outcome occurring in one group divided by the odds of the outcome occurring in another group.

Advantages and disadvantages of the cross-sectional studies are summarized in **Table 16.1**.

Data that Can be Obtained from Cross-sectional Studies

We can make the following estimates (**Table 13.2**):

- » Prevalence.
- » Odds ratio.
- » Logistic regression models.

Table 16.1. Advantages and Disadvantages of Cross-sectional Studies

Advantages	Disadvantages
<ul style="list-style-type: none"> • Good for examining rare outcomes or outcomes with long latency • Relatively quick to conduct • Relatively inexpensive • Requires comparatively few subjects • Existing records may be used • Multiple exposures or risk factors can be examined 	<ul style="list-style-type: none"> • Susceptible to recall bias or information bias • Difficult to validate information • Control of extraneous variables may be incomplete • Selection of an appropriate comparison group may be difficult • Rates of disease in exposed and unexposed patients cannot be determined

Key Terms

Define the following terms.

Cross-sectional studies

Odds ratio

Prevalence

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. State the features of a case-control study design.
2. List some important advantages to using the case-control study design.
3. List some important disadvantages to using the case-control study design.
4. List the estimates that can be obtained with a case-control study.
5. Draw a diagram of the case-control study design.
6. Complete **Table AL16.1** with the description, strengths, and weaknesses of observational analytic study designs.

Table AL16.1. Description, Strengths, and Weaknesses of Observational Analytic Study Designs

Study Design	Description	Strengths	Weaknesses
Case-control			
Cross-sectional			
Nested case-control			
Cohort			

7. As the hospital epidemiologist, you have been requested by the hospital administration to study the effects of administering antibiotics to patients at different time frames (2-hour intervals up to 24 hours) before they have surgery that involves opening the chest cavity. The study is aimed at reducing infections caused by surgery as well as reducing deaths. The study is to take place over the next 15 years.

- Design an appropriate study.
- Explain and justify the study design chosen.

8. You have been asked to study the effects of stress across the life span of people who have a close family member with HIV/AIDS. The study is to be from the time of diagnosis until the death of the family member.

- Design an appropriate study.
- Explain and justify the study design chosen.

9. Suppose 45 traffic accidents occur on a given road and you are interested in measuring whether the accidents are associated with rain showers. The “case” period is designated as the 24 hours preceding the accident and the “control” period is designated as 1 week prior to the case period. Among the accident cases, 7 experienced rain showers during the case and control periods; 16 experienced rain during the case period but not during the control period; 4 experienced no rain during the case period but rain during the control period; and 18 experienced no rain during either the case or control periods.

- Use an appropriate measure and describe whether an association exists between rain showers and traffic accidents.

10. Multiple choice questions.

1. A community assesses a random sample of its residents by telephone questionnaire. Obesity is strongly associated with diagnosed diabetes. This study design is best described as which one of the following:

- a) Case-control.
- b) Cohort.
- c) Cross-sectional.
- d) Experimental.

2. Based on a list of residents from election rolls, 2/3 of men in a large city are invited (including repeated educational urgings) and 1/3 of men are not invited to be screened by PSA blood test for prostate cancer. Over the next 10 years the two groups are compared as to the rate of death from prostate cancer. This study design is best described as which one of the following:

- a) Case-control.
- b) Cohort.
- c) Cross-sectional.
- d) Experimental.

3. Which of the following statements are true about an observational study?

- a) The researcher does not interfere with the study in any way.
- b) The researcher administers a treatment to the subjects in the study.
- c) The researcher is a subject in the study.
- d) None of these are correct.

4. You recall a conversation from your medical school days with one of your favorite anatomy professors. The professor observed that most students from his class who were good in anatomy tend to become radiologists. As a believer in science you decided to explore if there is any truth to this observation. Which study design is most suited to address the hypothesis that good anatomy students are most likely to become radiologists?
- Case-control.
 - Cohort.
 - Cross-sectional.
 - Randomized controlled trial.
5. What is the best design to study the prevalence of a disease?
- Case-control study.
 - Cohort study.
 - Cross-sectional study.
 - Randomized controlled trial.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Coggon D, Rose G, Barker DJP. Epidemiology for the uninitiated. 4ª Edition. London: BMJ Publishing Group Ltd; 2020. Available at: <https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated>.
- Setia MS. Methodology series module 3: Cross-sectional studies. Indian J Dermatol 2016;61:261-4.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.
- Johnson LL. Design of Observational Studies. In: Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3ª Edition. London: Elsevier Academic Press; 2012.
- Kesmodel US. Cross-sectional studies – what are they good for? Acta Obstet Gynecol Scand 2018; 97:388–393.
- Parfrey PS, Barret BJ. Clinical Epidemiology, Practice and Methods. 2ª Edition. New York: Springer; 2015.

Survival Analysis

Learning objectives for this chapter

- A. Recognize some vocabulary used in survival analysis and a few commonly used statistical methods for time to event data in medical research.
- B. Learn about censoring in survival analysis.
- C. Interpret a Kaplan-Meier graph.
- D. Identify applications with time to event outcomes.
- E. Define the term hazard ratio.
- F. Identify when a hazard ratio should be used.
- G. Understand how to interpret a hazard ratio.

With roots dating back to at least 1662 when John Graunt –a London merchant– published an extensive set of inferences based on mortality records, **Survival Analysis** is one of the oldest subfields of Statistics.

Basic life-table methods, including techniques for dealing with censored data, were discovered before 1700, and in the early eighteenth century, the old masters –de Moivre working on annuities, and Daniel Bernoulli studying competing risks for the analysis of smallpox inoculation– developed the modern foundations of the field.

Today, Survival Analysis models are important in Medicine, as well as in many more application areas such as Engineering, Insurance, and Marketing.

Survival analysis is a collection of statistical procedures for data analysis, for which the **outcome variable** of interest is the **time until an event occurs**.

Although the term **survival** is used, the event of interest is not limited to death or failure. **Other end points** can be used: recurrence of a supraventricular arrhythmia after ablation, pacemaker failure, relapse after leukemia treatment, readmission for congestive heart failure, and so on.

Death is the prime example of an outcome event used in survival analysis.

Survival time is defined as the time from some fixed starting point (time origin) to the onset of the event of interest.

- » In **controlled clinical trials**, the starting point is the actual time a participant enters the study, thus the starting point may vary for each participant.
- » In **Epidemiology**, the time origin may be birth, time of first exposure, or another point in time.

Key Features of Survival Data

Length of Follow-up

The length of follow-up time varies among participants. Patients entering later to the study would have a shorter follow-up than those entering earlier.

Outcome of Interest

By the end of the study, the event of interest is almost never observed in all subjects.

Applications of the Survival Analysis

The following are several examples of questions for which survival analysis may be applied:

- » How long does symptom improvement last after an epidural injection?
 - Time to the recurrence of back pain and recurrence vs. non-recurrence.
- » How long is the duration of the effect of antiemetic prophylaxis given to prevent nausea and vomiting resulting from the use of intravenous patient-controlled opioid analgesia?
 - Duration of nausea/vomiting prevention and manifestation vs. non-manifestation.
- » How long does it take for postoperative cognitive dysfunction caused by general anesthesia to occur?
 - Time to the occurrence of cognitive dysfunction and occurrence vs. non-occurrence.

Censoring

Ideal data for survival analysis are those yielded by cases in which the time of treatment is clearly established and all participants are followed up until they experience the event.

However, the observation period may end without occurrence of the event. The survival time is called **censored** if the event is **not observed** by the end of the study.

Right Censoring

It is the most common type of censoring in clinical studies. This indicates that **the period of observation (trial duration) ended before the event occurred**.

Other reasons for censoring include the **withdrawal of participants** from the study and **loss of contact with participants** who move out of the study area.

Independent Censoring

This indicates censoring for reasons **unrelated to the outcome for each participant** (i.e., the occurrence of the event of interest or not).

Survival Function

Survival function, denoted by $S(t)$, is defined as the probability of the outcome event **not occurring up to a specific point in time**, including the time point of observation (t). For example, if the event is “recurrence of back pain,” the survival function is the “probability of not having back pain” up to a specific time.

It gives the probability that the random variable T exceeds the specified time t .

The survival function equation is:

$$S(t) = P(T > t) = 1 - F(t)$$

Where $t = 0$ corresponds to a probability of 1.0 (i.e., 100% survival at the onset), and the point in time with 50% survival probability is the median survival time.

The ratio of the number of events occurring during the entire study period to the total number of observations is termed the **incidence rate**.

If the survival function is known from theory or empirical observations, then it can be used to analyze the survival experience of a population at various points in time.

The survival function is often expressed as a **Kaplan-Meier curve**.

The concept of a survival function is essential for the understanding of survival analysis.

Kaplan-Meier Estimator

In the Kaplan-Meier method, the incidence rate as a function of time is calculated by putting the observations in **ascending order** of time until occurrence.

The standard estimator of the survival function is called the **product limit estimator**.

It is obtained by taking the product of a sequence of conditional probabilities in order to create the **Kaplan-Meier curve**, an estimate of the true survival function (**Figure 17.1**).

Although no standard has yet been established, it is a general practice to:

- » Show **censored data** as **points or symbols**.
- » Show **decreases in the survival rate** (corresponding to the occurrences of the event) as **steps**.

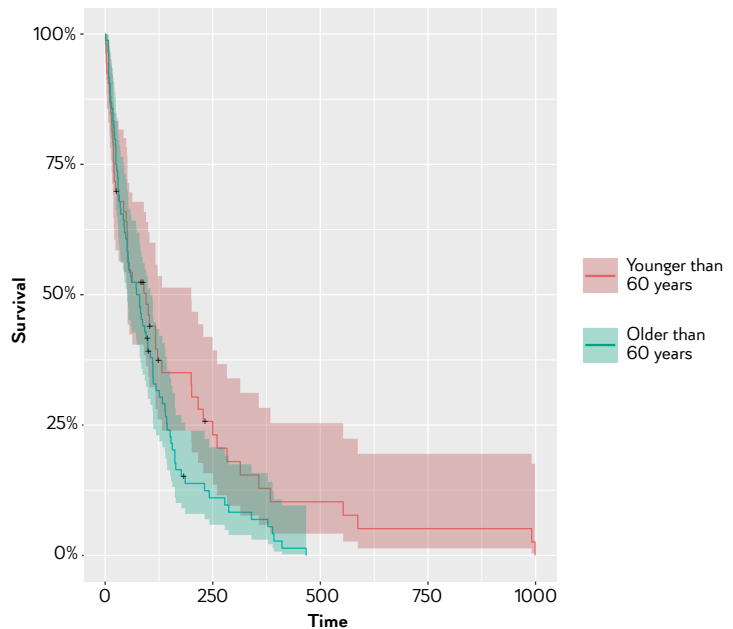


Figure 17.1. Kaplan-Meier curve of data from U.S. veterans subjected to a two-treatment, randomized trial for lung cancer, plotted by age. The veterans younger than 60 years are represented in red while those older than 60 are represented in green. Although the two curves appear to overlap in the first fifty days, younger patients clearly have a better chance of surviving more than a year.

Kaplan-Meier survival analysis is used to test for significant differences between survival curves and in median or mean survival time. Therefore, it fits well in studies focusing on survival time.

Cox Proportional Hazards Model

Time-to-Event Analysis

Clinical trials (**Chapter 28**) commonly record the **length** of time from study entry to a disease **endpoint** for a treatment and a control group. These data are commonly depicted with a Kaplan-Meier curve from which the **median** (time at which 50% of cases are resolved) and the **mean** (average resolution time) can be obtained.

Time-to-event analysis provides a method to **include patients who fail to complete the trial** or do not reach the study endpoint (**censored data**) by making **comparisons between** the number of **survivors in each group at multiple points in time**. Therefore, it is a potentially more powerful and informative method of analysis.

Excluding patients who are lost to follow-up, may introduce considerable **bias** because the data that these patients generate prior to their exit are important to the **power** and **validity** of the study.

Cox Proportional Hazards Regression Model

Also known as **Cox regression**, uses **regression analysis** (**Chapter 12**) to **process censored data**. This method can analyze any variables that may influence the occurrence of the event, and is defined as a **semi-parametric method**. It provides an estimate of the **hazard ratio** (HR) and its confidence interval.

A **Log-rank test** is another type of survival analysis that assesses if both curves differ significantly. Thus, it provides a **p-value**.

Hazard Ratio

A **hazard ratio** (HR) can be defined as an estimate of the ratio of the hazard rate in the treated versus the control group.

The **hazard rate** is the probability that, if the event in question has not already occurred, it will occur in the next time interval divided by the length of that interval.

The term **hazard** refers to the probability that an individual, under observation in a clinical trial at time t , has an event at that time.

Each **predictor variable** in a Cox regression model has a HR that tells you **how much the hazard increases in the relative sense** (that is, by what amount it's **multiplied**) when you increase the variable by exactly 1.0 unit. Therefore, a HR numerical value **depends on the units** in which the variable is expressed in the data.

Hazard ratios have also been used as a measure to describe the effect of an intervention on an outcome of interest over time.

The HR of two people are **independent of time**, and are valid only for **time-independent covariates**; the hazard functions for any two individuals at any point in time are **proportional**.

- » If an individual has a risk of death at some initial point in time that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.

Interpreting Hazard Ratios

The HR is a measure of the **magnitude of the difference** between the two curves in the Kaplan–Meier plot. The numerical value of the HR expresses the **relative hazard reduction** achieved by the study intervention compared to the hazard reduction in the control group.

The numerical value of a HR can be a fraction of 1.0 or it can be greater than 1.0:

- » A HR of **0.50** means that, at any particular time, **half** as many patients in the treatment group are experiencing an event compared to the control group.
- » A HR of exactly **1.0** means that at any particular time, event rates are the same in both groups.
- » A HR of **2.0** means that, at any particular time, **twice** as many patients in the treatment group are experiencing an event compared to the control group.

The HR is a **punctual estimate** and, therefore, its **confidence intervals (CI)** must be calculated.

- » The **narrower** the confidence interval, the **more precise the estimate**.
- » If the confidence interval **includes 1**, then the HR is **not significant**.

An Example from the Literature

Dupont et al. investigated the survival of patients with bronchiectasis according to age and use of long-term oxygen therapy. The Kaplan–Meier curves and results of the log rank tests shown in **Figure 17.2** indicate that there is a significant difference between the survival curves in each case.

The authors also applied Cox's regression and obtained the results given in **Table 17.1**. These results indicate that both age and long-term oxygen therapy have a significant effect on survival. The estimated risk ratio for age, for example, suggests that the risk for death for patients over the age of 65 years is 2.7 times greater than that for those below 65 years.

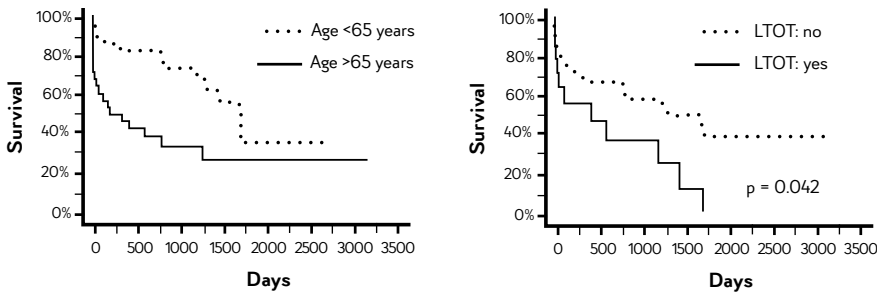


Figure 17.2. Kaplan-Meier curves showing survival rate of 48 patients with bronchiectasis after the first ICU admission for RF according to age on ICU admission (left) and use of LTOT before ICU admission (right).

Adapted from: Dupont M, Gacouin A, Lena H, Lavoue S, Brinchault G, Delaval P, Thomas R. Survival of patients with bronchiectasis after the first ICU stay for respiratory failure. *Chest*. 2004;125:1815-1820.

Table 17.1. Results of Cox's proportional hazards analysis for the patients with bronchiectasis

Explanatory variables	HR	95% confidence interval	p-value
Age (>65 years)	2.7	1.15–6.29	0.022
Long-term oxygen therapy (LTOT)	3.12	1.47–6.90	0.003

Key Terms

Define the following terms.

Censored data

Cox regression

Hazard

Hazard rate

Hazard ratio

Independent censoring

Kaplan-Meier curve

Length of follow-up

Product limit estimator

Right censoring

Survival analysis

Survival function

Survival time

Time-to-Event Analysis

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. State some examples for which survival analysis may be applied.

2. Based on Figure 17.1, estimate the median and mean for each of the groups of veterans subjected to a two-treatment randomized trial for lung cancer.

3. Multiple choice questions.

1. Kaplan–Meier statistics are most appropriate when:

- a) The time of follow-up is normally distributed.
- b) Many people drop out of the study.
- c) Most people have been followed for a long period of time.
- d) People have been followed for different periods of time.

2. If Kaplan–Meier statistics are used, information of the event should be collected from participants:

- a) At regular 6-month intervals.
- b) As often as possible.
- c) At planned medical check-ups.
- d) At planned home visits.

3. A censored observation in the data occurs:

- a) Before the follow-up data are collected.
- b) When the data are not entered into the database.
- c) If a participant misses a study visit.
- d) When a person has withdrawn from the study.

4. If the y-axis of a Kaplan–Meier curve is shortened this will:

- a) Visually magnify differences between study groups.
- b) Visually minimize differences between study groups.
- c) Visually make no difference.
- d) Visually make the figure easier to read.

5. The Cancer Prevention Study II (Harris et al. 2004) followed a cohort of 364 239 men and 576 535 women for a period of 6 years to determine rates of death from cancer of the trachea, bronchus or lung. The authors considered the association between mortality rates and the tar level of the cigarettes smoked by the subset of men in the study who were current smokers in 1982. Compared to men who smoked cigarettes with a tar content of 15–21 mg, the mortality hazard ratios among those who smoked cigarettes with a tar content of 0–7 mg, 8–14 mg and ≥ 22 mg were 1.17 (95% confidence interval 0.95 to 1.45), 1.02 (0.90 to 1.16) and 1.44 (1.20 to 1.73), respectively. Which one of the following statements is true?

- a) There is no evidence from this study that the tar content of the cigarettes smoked is associated with an increased mortality rate from cancer of the trachea, bronchus or lung.
- b) Men who smoked cigarettes with a tar content of ≥ 22 mg had a significantly increased risk of mortality from cancer of the trachea, bronchus or lung compared to men who smoked cigarettes with a tar content of 0–7 mg.
- c) The authors would have been able to more usefully estimate the association between tar content and mortality risk if they had included tar content as a continuous covariate in their analysis.
- d) The reference group for the analysis was men who did not smoke cigarettes in 1982.

e) Had the authors changed the reference group for the analysis to men who smoked cigarettes with a tar content of ≥ 22 mg, the relative hazard estimate for those smoking cigarettes with a tar content of 15–21 mg would have been less than 1.

6. T Which one of the following statements about survival analysis is true?

- a) What is of primary importance in a survival analysis is whether or not the individual reaches the endpoint (e.g. death).
- b) Survival times are right-censored where followup does not begin until after the baseline date.
- c) Informative censoring means that full information is available on why and when an individual's followup is censored.
- d) The relative hazard is assumed to be constant in a Cox proportional hazards model.
- e) The log-rank test in a Kaplan–Meier survival analysis is a para metric test, comparing the survival experience in two or more groups, which assumes that the logarithm of the ranked data are Normally distributed.

7. In a survival curve the y-axis represents:

- a) The time taken for participants to experience the event.
- b) The proportion of participants yet to experience the event.
- c) The probability of experiencing the event.
- d) None of the above.

8. The definition of the hazard function is:

- a) The rate of survival at each time-point.
- b) The rate of the event of interest at a specified time-point.
- c) The dangers associated with the conditions in the study.
- d) All of the above.

9. When a person does not experience the event within the time-frame for the study they are called:

- a) Surplus to requirements.
- b) An invalid case.
- c) An outlier.
- d) Right censored.

10. Which of the following represents the best definition of censored cases?

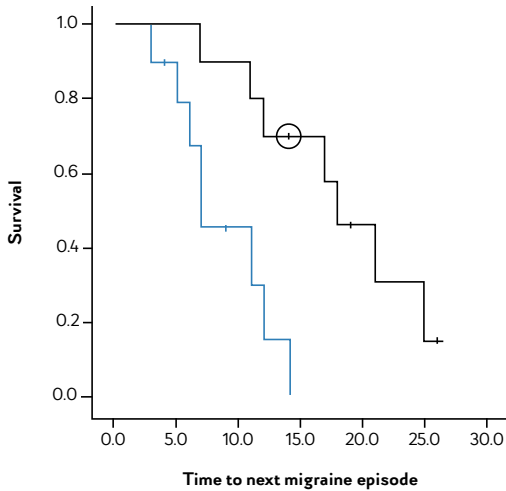
- a) People who do not want to join the study.
- b) Participants who drop out of the study and/or have not experienced the event of interest.
- c) Participants who don't read the instructions carefully enough.
- d) Both b) and c) above.

11. The beginning of the time-period for a survival analysis is often called:

- a) The start of the study.
- b) The time of randomization.
- c) Time zero.
- d) Let's get started.

12. Take a look at the survival curve on the next page. What can you say about the person indicated in the graph by the circle around the line?

- a) The person had a migraine in week 14.
- b) The person did not have a migraine in week 14 but had it at a later time.
- c) The person did not have a migraine at all within the time-frame of the study.
- d) The person was lost to the study in week 14.



13. Referring back to the graph in question 12, what term is given to the person highlighted?

- They are unimportant.
- They are censored.
- They are right censored.
- They are a hazard.

14. How might we calculate the probability of a participant surviving until the fifth day in the study?

- It is the probability of not experiencing the event divided by the probability of experiencing the event.
- It is the probability of surviving until the fourth day multiplied by the probability of surviving the fifth day.
- It is the probability of surviving the first day, times the probability of surviving the second day, times the probability of surviving the third day, times the probability of surviving the fourth day, times the probability of surviving the fifth day.
- Both b) and c) above.

15. Is the probability that, if the event in question has not already occurred, it will occur in the next time interval divided by the length of that interval:

- Hazard.
- Hazard ratio.
- Hazard rate.
- Hazard reduction.

Bibliography and Suggested Reading

- Barracough H, Simms L, Govindan R. Biostatistics Primer: What a Clinician Ought to Know: Hazard Ratios. *J Thorac Oncol.* 2011 Jun;6(6):978-82.
- Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. *Critical Care.* 2004; 8:389-394.
- Brody T. Biostatistics—Part I. In: Brody T. Clinical trials. Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. 2o Edition. Cambridge: ELSEVIER; 2016.
- Dupont M, Gacouin A, Lena H, Lavoue S, Brinchault G, Delaval P, Thomas R. Survival of patients with bronchiectasis after the first ICU stay for respiratory failure. *Chest.* 2004;125:1815-1820.
- Hoffman JIE. Survival Analysis. In: Hoffman JIE. Basic Biostatistics for Medical and Biomedical Practitioners. Waltham: Elsevier Academic Press; 2019.
- In J, Lee DK. Survival analysis: Part I — analysis of time-to-event. *Korean J Anesthesiol.* 2018;71(3): 182-191.
- Johnson LL, Shih JH. An Introduction to Survival Analysis. In: Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3° Edition. London: Elsevier Academic Press; 2012.
- Parfrey PS, Barret BJ. Clinical Epidemiology. Practice and Methods. 2° Edition. New York: Springer; 2015.
- Qin J. A Review of Survival Analysis. In: Qin J. Biased Sampling, Over-identified Parameter Problems and Beyond. Bethesda: Springer; 2017.
- Sedgwick P. Hazards and hazard ratios. *BMJ.* 2012; 345: e5980.
- Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. *Perspect Clin Res* 2011;2:145-8.
- Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. *Antimicrob Agents Chemother.* 2004; 48: 2787-2792.

Disease Occurrence, Risk, Association, Importance, and Implication

Learning objectives for this chapter

- A. Understand the concept of epidemiological measures.
- B. Learn the classification of epidemiological measures.

Epidemiology studies the frequency of health events and their distribution patterns according to the characteristics of the population, the regions and moments in time. It also analyzes the determinants and factors that generate the observed panorama and, based on this, proposes and evaluates the corresponding intervention measures, either to avoid new cases or to control the existing ones and minimize the sequelae left by the pathology.

In order to achieve this task, Epidemiology has various **epidemiological measurements** that can be classified into three categories (**Table 18.1**):

1. Measures of **disease occurrence and risk**.
2. Effect measures for the **association between a disease and an exposure**.
3. Measures of **importance or implication**.

Table 18.1. Examples of Different Effect Measures in Epidemiological Studies

Measures of Occurrence and Risk	Effect Measures for Association	Measures of Importance or Implication
Incidence Prevalence	Relative risk (RR) Odds ratio (OR) Correlation	Excess risk Attributable risk (AR) Population attributable risk (PAR)

Measures of Disease Occurrence and Risk: Incidence and Prevalence

Are used to describe **causal relationships** and in descriptive analyses of the **evolution of disease occurrence or mortality over time**.

The concept of **disease risk** is often used to describe the association between an exposure and the probability of having a given disease now or developing that disease later in life.

To determine the **risk** associated with a given exposure, the population studied must be divided into **two or more groups** (e.g., an exposed group and an unexposed group), or by level of exposure. The importance of the exposure can be assessed when the number of exposed and unexposed subjects that either have a disease or developed a disease during a defined time period has been determined. With this information, the **risk associated** with the exposure can be estimated. The association between the occurrence of a disease and a given exposure can be summarized in a contingency table (**Table 18.2**).

Incidence

Incidence is defined as the number of individuals newly diagnosed with disease in a defined time period.

Incidence rate is the incidence divided by the length of this time period.

The following **formula** can be used to calculate the incidence:

$$\text{Incidence} = \frac{\text{Number of new cases during a defined time period}}{\text{total population in risk at the beginning of the study}}$$

Prevalence

Prevalence is defined as the proportion of a population with a given disease at a set point in time.

Table 18.2. Example of a 2 x 2 Contingency Table: Cell Counts for the Association Between Disease and Exposure

Exposure	Disease		Total
	Yes	No	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	n

The following **formula** can be used to calculate the prevalence:

Prevalence = Number of cases with the disease at a set point in time/total population in the same group and at the same set point in time

Duration of disease, together with the incidence, rate will determine the prevalence.

There are two measures of prevalence:

- **Point prevalence:** is the probability an individual will have a given disease at a set point in time (t).
- **Period prevalence:** is an expression of a probability that an individual has been affected by a given disease during a defined time period.

Effect Measures: RR and OR

The effect measure yields an **estimate of the strength of the association** between two variables (e.g., the association between a possible causal factor and a disease).

Effect measures are based on comparisons of disease occurrence in groups with various levels of exposure to a certain variable.

The effect measure may be expressed as the **risk or the probability of disease** occurrence during a defined time period.

Relative Risk (RR) and **Odds Ratio (OR)** are fully described in **Chapter 19**.

Measures of Importance or Implication: AR, PAR, and Excess Risk

They express the **impact of a certain disease** or **exposure on a population**.

These measures are of **Public Health interest**, meaning that they also help to assess the possible effect of preventive efforts.

Attributable Risk (AR) and **Population Attributable Risk (PAR)** are fully described in **Chapter 21**.

Key Terms

Define the following terms.

Attributable risk

Epidemiological measures

Incidence

Incidence rate

Odds ratio

Period prevalence

Point prevalence

Population attributable risk

Prevalence

Relative risk

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Explain the difference between incidence and prevalence of a disease.

2. Frequently, family medicine services are interested in knowing the prevalence of patients with diabetes in the community where they provide the medical service. To obtain this information, the epidemiologist orders his team to visit and record the entire population covered by his unit. As a result of his research, he finds that 228 diabetic subjects were identified in a population of 4 550 inhabitants.

- Calculate the prevalence of diabetes in this population.

3. A cohort of 12 subjects is taken and observed for five years (without their elements being lost for the study and without ceasing to be at risk during the study period). Of these, five develop the disease.

- Calculate the incidence for this 5 years.
- If an individual is taken from those observed at the start of the study, what is the probability that the disease will develop within five years?

4. In a survey of 1,150 women who gave birth in Maine in 2000, a total of 468 reported taking a multivitamin at least 4 times a week during the month before becoming pregnant.

- Calculate the prevalence of frequent multivitamin use in this group.

5. You study a population of 20 persons with 10 new cases of an illness over about 15 months. Before October 1, six people became ill; 2 of them died before April 1. Between October 1 and September 30, four more persons became ill. Six more persons died after April 1.

- Calculate the incidence rate from October 1, 2004, to September 30, 2005, using the midpoint population (population alive on April 1, 2005) as the denominator.
- Calculate the point prevalence on April 1, 2005. Point prevalence is the number of persons ill on the date divided by the population on that date. On April 1, seven persons (persons 1, 4, 5, 7, 9, and 10) were ill.
- Calculate the period prevalence from October 1, 2004, to September 30, 2005. The numerator of period prevalence includes anyone who was ill any time during the period. In Figure 3.1, the first 10 persons were all ill at some time during the period.

6. Lung cancer is the leading cause of cancer death for both men and women in the United States. More people die of lung cancer than of colon, breast, and prostate cancers combined.

Lung cancer is fairly rare in people under the age of 40. The average age of people found to have lung cancer is 60. In 2004 there will be about 173,770 new cases of lung cancer in the United States. About 160,440 people will die of this disease. The population of the United States in 2004 is 292,287,454.

- Calculate the annual incidence rate of lung cancer in the US in 2004.

7. During a given year there were 30 new cases of disease X in population A and 3 new cases in population B.

- Based on this finding, is it accurate to say that the rate of disease is higher in population A? Why or why not?

8. In a class of 26 people, none had upper respiratory symptoms at the beginning of the semester. One week later, 3 students reported having upper respiratory symptoms. One week after that (beginning of week 2), 1 student had recovered but there were 2 new cases.

- Calculate the prevalence of upper respiratory symptoms at the beginning of week 2.
- Calculate the risk (incidence proportion) of developing upper respiratory symptoms.
- Calculate the rate of upper respiratory symptoms per 100 student-weeks.

9. The incidence rate of a disease is 50 per 100,000 person-years. The average duration of the disease is 2 years, after which patients fully recovers.

- Estimate the prevalence of the disease in the population assuming the population is stationary and disease occurrence is in a steady state.

10. 200 healthy men are followed for the occurrence of prostate cancer. After 5 years, 30 cases occur.

- Calculate the incidence rate of prostate cancer in this cohort with and without an actuarial correction.

Bibliography and Suggested Reading

- Arnett DK, Claas SA. Introduction to Epidemiology. In: Robertson D, Williams GH. Clinical and Translational Science: Principles of Human Research. London: ELSEVIER; 2009.
- Pearl J. Causality: models, reasoning, and inference. 2nd Edition. Cambridge: Cambridge University Press; 2009.
- Perez L, Künzli N. From measures of effects to measures of potential impact. *Int J Public Health*. 2009; 54:45–48.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.

Odds Ratio and Relative Risk: As Simple as It Can Get

Learning objectives for this chapter

- A. Define the terms relative risk and odds ratio.
- B. Identify when relative risk and odds ratio should be used.
- C. Calculate a relative risk and odds ratio from a 2 x 2 table.
- D. Understand how to interpret confidence intervals around a relative risk or odds ratio.

Researchers are often interested in evaluating the **association** between an **exposure and an outcome**. In other words, they are interested in knowing whether the presence of a risk factor, or performing an intervention, alters the risk of an outcome as compared to the absence of the risk factor or the intervention. In analytical studies, it is not only interesting to know if this association exists, but also the **magnitude of that association**. This is achieved by comparing the frequency of the event of interest in a group exposed to the study factor with that of an unexposed group.

Risk and Odds, Is There a Difference?

- » “**Risk**” refers to the probability of occurrence of an event. Statistically, risk refers to chance of the outcome of interest divided by all possible outcomes.
- » “**Odds**” refers to the probability of occurrence of an event divided by the probability of the event not occurring.

At first glance, you may think that both concepts seem similar and interchangeable. Nevertheless, there are important differences that dictate where the use of either of these is appropriate.

Let’s discuss the following example in order to fully understand the differences between risk and odds.

You are reading a randomized clinical trial comparing endoscopic sclerotherapy ($n = 65$) versus band ligation ($n = 64$) for the treatment of bleeding esophageal varices (**Table 19.1**).

Table 19.1. A Randomized Clinical Trial of Sclerotherapy vs. Ligation for Esophageal Varices (hypothetical data)

Intervention	Outcome		Total
	Death	Survival	
Ligation	18	46	64
Sclerotherapy	29	36	65
Total	47	82	129

Based on the data in the **Table 19.1**, we can conclude:

- » The **overall risk of death** = $47/129$ ([number of deaths]/[all outcomes i.e., all deaths + survivors]) = **0.36**.
- » The **overall odds of death** = $47/82$ ([number of deaths]/[number of no deaths, i.e., survivors]) = **0.57**.
- » The **risk of death in the ligation group** was $18/64$ (28% or 0.28).
- » The **risk of death in the sclerotherapy group** was $29/65$ (44% or 0.44).
- » The **odds of death in the two groups** was $18/46$ (0.39) and $29/36$ (0.81), respectively.

From the data in the previous example, the chances of death appear markedly **different** when expressed as risks and odds.

Now, let's consider another scenario based on **Table 19.2**:

- » As "a" decreases with respect to "b" (probability of outcome becomes less), the odds and risk are similar.
- » For rare events (i.e., if "a" is small and "a + b" approaches "b"), $a/(a + b) \approx a/b$ and risk approximates odds.
- » Therefore, though "odds" does not represent true risk, its value is close to risk when the event rates are low (typically <10%).

Table 19.2. Example of a 2 x 2 Contingency Table for the Calculation of Association Measures

	Disease	No disease	Total
Exposure	a	b	a + b
No Exposure	c	d	c + d
Total	a + c	b + d	a + b + c + d

Some studies use **relative risks (RR)** to describe results; others use **odds ratios (OR)** to do so. Both association measures are calculated using a **2 × 2 contingency table** like the one shown in **Table 19.2**.

Relative Risk

The relative risk (RR) of an event estimates the **magnitude of an association** and indicates the **number of times that an event (disease) is more likely to develop in a group exposed to a risk variable compared to a group unexposed to the same risk variable**.

The RR is the division between the incidence in the exposed group (I_e) and the incidence among the non-exposed group (I_o).

- » A RR of **1.0** indicates that there is **no relationship between the study factor and the disease** because the risk of the event is identical in the exposed and not exposed groups.
- » If the RR is **greater than 1.0**, it indicates a **positive association between the study factor and the disease** because the risk is increased in the exposed sample.
- » If the RR is **less than 1.0**, it indicates a **negative association between the study factor and the disease** because the risk is lower in the exposed sample.

The RR obtained in a study is a **punctual estimate** and, therefore, its **confidence intervals (CI)** must be calculated.

- » If the 95% CI **does not include the RR = 1 value**, there is a **statistically significant association** between the study factor and the outcome.
 - For example: if the RR is 1.70, and the CI is 0.90–2.50, the elevation in risk is not statistically significant because the value 1.00 (no difference in risk) lies within the confidence interval.

Given the fact that RR is based upon the **incidence** of a given event in which we already know the participants' exposure status, it is only appropriate to **use it for prospective cohort studies**.

Bridge to Cohort Studies



Prospective cohort studies are carried out from the present time into the future. The researcher divides the subjects into exposed and not exposed to a possible risk factor, and follows them for a while.

How to Interpret a RR?

Let's say that we have an RR of 0.3. In "plain English" this can be expressed in many ways:

- » The risk is lowered to less than one-third.
- » The risk is reduced to 30%.
- » The risk is lowered by more than two-thirds.
- » The risk is reduced by 70%.

Odds Ratio

Odds means the **ratio between the probability that an event will occur and the probability that it will not occur.**

If the probability that a person with a disease is exposed to the study factor is 0.75, then the odds of the exposure will be calculated by dividing this value by the probability of not being exposed ($0.75 / [1 - 0.75] = 3$).

- » Thus, the odds of rolling 6 with a die are 1 to 5 (i.e., 0.20); this contrasts with the risk of rolling 6, which is $1/6$ (i.e., 0.17).
- » Similarly, the odds of tossing heads with a coin are 1 to 1 (or “50-50,” or 1.00), whereas the risk of tossing heads is $1/2$ or 0.5.

Odds ratio is a comparison of the odds of an event after exposure to a risk factor in the case group (a/c) with the odds of that event in the control group (b/d).

The OR is **numerically different** from the RR, even though both seek to compare the same risk between the same groups.

- » An OR of **1.0** indicates that there is **no increase or decrease in risk.**
- » If the OR is **greater than 1.0**, it indicates that **the exposure to the risk variable increases the risk** of the event.
- » When the OR is **less than 1.0**, it indicates that **the exposure to the risk variable reduces the risk** of the event.

The OR obtained in a study is also a **punctual estimate**. Therefore, its **confidence intervals (CI)** must also be calculated.

- » If the 95% CI for the OR **includes 1.00**, the OR is **not statistically significant.**

Case-control studies

A **case-control study** compares patients who have a disease or outcome of interest (cases) with patients who do not have it (controls), and looks retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease.



OR are **always obtained from case-control studies** because in this type of studies incidence cannot be calculated because the study population is selected from individuals who already have developed the disease.

How to Interpret an OR?

The only interpretation for an OR is “**times for**” and a **1 must be subtracted to the result** in order to be a **valid interpretation**. ORs **should not be interpreted as percentages** because the effect observed would be **overestimated**.

In order to fully understand how to interpret an OR, let’s analyze the results obtained by Pesch, et al. about smoking and lung cancer.

In their article, they concluded that male smokers with an average daily dose of >30 cigarettes had:

- » An OR of 103.5 (95% CI 74.8-143.2) for squamous cell carcinoma.
- » An OR of 111.3 (95% CI 69.8-177.5) for small cell lung cancer.
- » And an OR of 21.9 (95% CI 16.6-29.0) for adenocarcinoma.

That means that:

- » Smoking increases 102.5 times the risk of developing squamous cell carcinoma.
- » Smoking increases 110.3 times the risk of developing small cell lung cancer.
- » Smoking increases 20.9 times the risk of developing adenocarcinoma.

Key Terms

Define the following terms.

2 × 2 contingency table

Odds ratio

Relative risk

Confidence interval

Overall odds of death

Risk

Odds

Overall risk of death

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Analyze the following abstract from a paper by Illi, et al.

Objective: To investigate the association between early childhood infections and subsequent development of asthma.

Design: Longitudinal birth cohort study.

Setting: Five children's hospitals in five German cities.

Participants: 1314 children born in 1990 followed from birth to the age of 7 years.

Main outcome measures: Asthma and asthmatic symptoms assessed longitudinally by parental questionnaires; atopic sensitization assessed longitudinally by determination of IgE concentrations to various allergens; bronchial hyperreactivity assessed by bronchial histamine challenge at age 7 years.

Results: Compared with children with 1 episode of runny nose before the age of 1 year, those with 2 episodes were less likely to have a doctor's diagnosis of asthma at 7 years old (odds ratio 0.52 (95% confidence interval 0.29 to 0.92)) or to have wheeze at 7 years old (0.60 (0.38 to 0.94)), and were less

likely to be atopic before the age of 5 years. Similarly, having 1 viral infection of the herpes type in the first 3 years of life was inversely associated with asthma at age 7 (odds ratio 0.48 (0.26 to 0.89)). Repeated lower respiratory tract infections in the first 3 years of life showed a positive association with wheeze up to the age of 7 years (odds ratio 3.37 (1.92 to 5.92) for 4 infections v 3 infections).

Conclusion: Repeated viral infections other than lower respiratory tract infections early in life may reduce the risk of developing asthma up to school age.

- What is meant by odds ratio 0.52 for runny nose and asthma and what does it tell us?
- What is meant by 95% confidence interval 0.29 to 0.92 and what further information does this provide?
- What is meant by odds ratio 3.37 (1.92 to 5.92) for lower respiratory tract infections and wheeze?
- On a less statistical point, what is wrong with the way the conclusion is phrased?

2. Analyze the following abstract from a paper by Towner, et al.

Objective: To apply a measure of exposure to injury risk for schoolchildren aged 11-14 across a population and to examine how risk factors vary with sex, age, and affluence.

Design: Self completion questionnaire survey administered in schools in May 1990. Setting : 24 schools in Newcastle upon Tyne.

Subjects: 5334 pupils aged 11-14, of whom 4637 (87%) completed the questionnaire.

Results: Boys were exposed to greater risk than girls in journeys to places to play outdoors: they took longer trips and were more likely to ride bicycles (relative risk 5.30 (95% confidence interval 4.23 to 6.64) and less likely to travel by public transport or car. Younger pupils (aged 11-12) were less exposed to traffic during journeys to and from school: their journeys were shorter, they were less likely to walk (trip to school, relative risk 0.88 (0.83 to 0.94), and they were more likely to travel by car (trip to school, relative risk 1.33 (1.13 to 1.56)) or school bus (1.33 (1.10 to 1.62)). Poorer children were exposed to greater risk than affluent children (from families that owned a car and a telephone): they were less likely to travel to school by car (relative risk 0.26 (0.20 to 0.33)) or to be accompanied by an adult (0.39 (0.32 to 0.48)).

Conclusion: Injury risk data can provide useful information on child injury prevention and can be used to identify priorities and target resources

for injury prevention on a citywide scale or for an individual school.

- What is meant by 'relative risk 5.30'?
- What would the relative risk of riding bicycles be if boys and girls were equally likely to report riding bikes?
- Is there good evidence that younger children were less likely to walk to school than were older children?
- Is there good evidence that children from poorer families were more likely to be accompanied by an adult than were children from more affluent families?
- Why must fewer than 20% of girls have reported riding bicycles? (N.B. the question was actually about their last journey to play, not whether they ever use them.)

3. In an outbreak of tuberculosis among prison inmates in South Carolina in 1999, 28 of 157 inmates residing on the East wing of the dormitory developed tuberculosis, compared with 4 of 137 inmates residing on the West wing.

- Summarize these data in a 2x2 contingency table.
- In this example, the exposure is the dormitory wing and the outcome is tuberculosis. Calculate the risk ratio.
- How is this RR interpreted?

4. In an outbreak of varicella (chickenpox) in Oregon in 2002, varicella was diagnosed in 18 of 152 vaccinated children compared with 3 of 7 unvaccinated children.

- Summarize the data in a 2x2 contingency table.
- Calculate the risk ratio.
- How is this RR interpreted?

5. Calculate the odds ratio for the tuberculosis data in Exercise 3.

- Would you say that the odds ratio is an accurate approximation of the risk ratio?

6. True or False: You can use a 95% confidence for the odds ratio to determine statistical significance at $\alpha = 0.05$.

7. True or False: You can use a 95% confidence for the odds ratio to determine statistical significance at $\alpha = 0.01$.

8. Results from two case-control studies on cell phone use and brain cancer are summarized below. Review each summary and discuss whether the study in question supports or does not support the theory that recent use of hand-held cellular telephones causes brain tumors. Explain your reasoning in each instance.

- In 2001, a case-control study by Inskip et al. examined the use of cellular telephones between 1994 and 1998 in 782 cases with various forms of intracranial tumors and 799 controls admitted to the same hospitals for a variety of nonmalignant conditions. Subjects were considered exposed if they reported use of a cellular telephone for more than 100 hours. The odds ratio (OR) for glioma was 0.9 (95 percent confidence interval 0.5 to 1.6), the OR for meningioma was 0.7 (95 percent confidence interval 0.3 to 1.7), the OR for acoustic neuroma 1.4 (95 percent confidence interval 0.6 to 3.5), and the OR for all tumor types combined: 1.0 (95 percent confidence interval 0.6 to 1.5)
- In 2000, a case-control study by Muskat, et al. conducted between 1994 and 1998 used a structured questionnaire to quantify the statistical relation between cell phone use and primary brain cancer in 469 cases and 422 matched controls. The results of the study stated “The median monthly hours of use were [sic] 2.5 for cases and 2.2 for controls. Compared with patients who never used handheld cellular telephones, the multivariate odds ratio (OR) associated with regular past or current use was 0.85 (95% confidence interval [CI], 0.6-1.2). The OR for infrequent users (<0.72 h/mo) was 1.0 (95% CI, 0.5-2.0) and for frequent users (>10.1 h/mo) was 0.7 (95% CI, 0.3-1.4). The mean duration of use was 2.8 years for cases and 2.7 years for controls . . . The OR

was less than 1.0 for all histologic categories of brain cancer except for uncommon neuroepitheliomatous cancers (OR, 2.1; 95% CI, 0.9-4.7).”

9. A fictitious study was conducted to determine the effect of oral contraceptive (OC) use on heart disease risk in 40- to 44-year old women. This study found 13 new cases among 5000 OC users over 3-years of follow-up. In contrast, among 10,000 non-users, 7 developed a first myocardial infarct.

- Summarize the data in a 2x2 contingency table.
- Calculate the risk ratio, and interpret your results.

10. Multiple choice questions.

1. The odds ratio is:

- The ratio of the probability of an event not happening to the probability of the event happening.
- The probability of an event occurring.
- The ratio of the odds after a unit change in the predictor to the original odds.
- The ratio of the probability of an event happening to the probability of the event not happening.

2. In a cohort study, the risk ratio of developing diabetes was 0.86 when comparing consumers of tea (the exposed) to those who did not drink tea (the unexposed). Which one statement is correct?

- The tea drinkers have lower risk of developing diabetes.
- The tea drinkers have higher risk of developing diabetes.
- Based on the information given we cannot tell if the observed difference in disease risk is the result of chance.
- The risk ratio is close to the value one, so there is no difference in disease risk between the two groups.

3. Relative risk:

- a) Shows the relationship between a factor assumed to influence the occurrence of disease, and the disease.
- b) Is the ratio of the risk of disease for those exposed and those not exposed to that risk factor.
- c) Cannot be greater than 1.
- d) Is always expressed as a percentage.

4. A cohort study of smoking and lung cancer was conducted in a small island population. There were a total of 1,000 people in the study, and the study was conducted over a ten year period. Four hundred were smokers and 600 were not. Of the smokers, fifty developed lung cancer. Of the non-smokers, 10 developed lung cancer. In order to measure the strength of association between smoking and lung cancer in this population, which measure of exposure-disease association would you use?

- a) Risk ratio.
- b) Risk difference.
- c) Odds ratio.
- d) Attack rate.

5. In the previous cohort study examining the association between smoking and lung cancer, suppose you obtain a measure of exposure-disease association = 17. How would you interpret this in words?

- a) There were 17 more cases of lung cancer in the smokers.
- b) Smokers had 17% more lung cancers compared to non-smokers.
- c) Smokers had 17 times the risk of lung cancer compared to non-smokers.
- d) 17% of the lung cancers in smokers were due to smoking.

6. A group of patients with lung cancer is matched to a group of patients without lung cancer. Their smoking habits over the course of their lives are compared. On the basis of this information, researchers compute the odds of smoking among patients with lung cancer compared to the odds of smoking among those without lung cancer. This is an example of a:

- a) Case-control study.
- b) Cohort study.
- c) Cross-sectional study.
- d) Longitudinal study.

7. A study is performed in which mothers of babies born with neural tube defects are questioned about their acetaminophen consumption during the first trimester of pregnancy. At the same time, mothers of babies born without neural tube defect are also questioned about their consumption of acetaminophen during the first trimester. Which of the following measures of association is most likely to be reported by investigators?

- a) Prevalence.
- b) Relative risk.
- c) Odds ratio.
- d) Hazard ratio.

8. An observational study in diabetics assesses the role of an increased plasma fibrinogen level on the risk of cardiac events. 130 diabetic patients are followed for 5 years to assess the development of acute coronary syndrome. In the group of 60 patients with a normal baseline plasma fibrinogen level, 20 develop acute coronary syndrome and 40 do not. In the group of 70 patients with a high baseline plasma fibrinogen level, 40 develop acute coronary syndrome and 30 do not. Which of the following is the best estimate of relative risk in patients with a high baseline plasma fibrinogen level compared to patients with a normal baseline plasma fibrinogen level?

- a) $(40/30)/(20/40)$.
- b) $(40*40)/(20*30)$.
- c) $(40*70)/(20*60)$.
- d) $(40/70)/(20/60)$.

9. A trial comparing a drug to placebo stated OR 0.5 95%CI 0.4–0.6. What would it mean?

- a) The odds of death in the drug arm are 50% less than in the placebo arm with the true population effect between 20% and 80%.
- b) The odds of death in the drug arm are 50% less than in the placebo arm with the true population effect between 60% and 40%.
- c) The odds of death in the SuperStatin arm are 50% less than in the placebo arm with the true population effect between 60% and up to 10% worse.

10. If the trial from question 9 stated OR 0.5 95%CI 0.4–0.6 $p < 0.01$. What would it mean?

- a) The odds of death in the drug arm are 50% less than in the placebo arm with the true population effect between 60% and 40%. This result was statistically significant.
- b) The odds of death in the drug arm are 50% less than in the placebo arm with the true population effect between 60% and 40%. This result was not statistically significant.
- c) The odds of death in the drug arm are 50% less than in the placebo arm with the true population effect between 60% and 40%. This result was equivocal.

Bibliography and Suggested Reading

- Andrade C. Understanding Relative Risk, Odds Ratio, and Related Terms. *J Clin Psychiatry*. 2015;76(7):e857–e861.
- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Illi S, von Mutius E, Lau S, Bergmann R, Niggemann B, Sommerfeld C, Wahn U. Early childhood infectious diseases and the development of asthma up to school age: a birth cohort study. *British Medical Journal*. 2001;322:390–395.
- Inskip PD, Tarone RE, Hatch EE, Wilcosky TC, et al. Cellular-Telephone Use and Brain Tumors. *N Engl J Med*. 2001; 344:79–86.
- Last A, Wilson S. Relative risks and odds ratios: What's the difference? *J Fam Pract*. 2004 Feb;53(2):108.
- McHugh ML. The odds ratio: calculation, usage, and interpretation. *Biochemia Medica*. 2009;19(2):120–6.
- Muscat JE, Malkin MG, Thompson S, et al. Handheld Cellular Telephone Use and Risk of Brain Cancer. *JAMA*. 2000;284(23):3001–3007.
- Parfrey PS, Barret BJ. *Clinical Epidemiology, Practice and Methods*. 2° Edition. New York: Springer; 2015.
- Pesch B, et al. Cigarette smoking and lung cancer – relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int J Cancer*. 2012; 131(5):1210–1219.
- Ranganathan P, Aggarwal R, Pramesh CS. Common pitfalls in statistical analysis: Odds versus risk. *Perspect Clin Res*. 2015;6:222–4.
- Schmidt CO, Kohlmann T. When to use the odds ratio or the relative risk? *Int J Public Health*. 2008;53:165–167.
- Sedgwick P. Relative risks versus odds ratios. *BMJ*. 2014;348:g1407.
- Towner EML, Jarvis SN, Walsh SSM, Aynsley-Green A. Measuring exposure to injury risk in schoolchildren aged 11-14. *British Medical Journal*. 1994;308:449–452.

Confounding

Learning objectives for this chapter

- A. Define confounding and distinguish it from bias and chance error.
- B. Explain the three key properties of a confounder.
- C. Diagram the relationship of a confounder with exposure and outcome.
- D. Understand that there are methods to adjust for confounding.
- E. Understand the Simpson's Paradox and apply it to clinical practice.

Confounding is an important concept in Epidemiology because if present, it can cause an **over- or under- estimate of the observed association** between an exposure and an outcome. The distortion introduced by a confounding factor can be large, and it can even change the apparent direction of an effect. However, it can be **adjusted for** in the statistical analysis.

What is Confounding?

Confounding can be defined as the **distortion of the association between an exposure and a health outcome caused by an extraneous, third variable called “confounder”**.

This confounder is **associated with both** the factor of interest and the outcome, and affects the variables under study. A graphical model of confounders is shown in **Figure 20.1**.

Confounding may be present in **any study design** (i.e., cohort, case-control, observational, ecological), primarily because **it's not a result of the study design**. However, of all study designs, **ecological studies** are the most susceptible to confounding.

Confounding variables are variables that are not central to your study but which may be responsible for the effect that you are interested in.

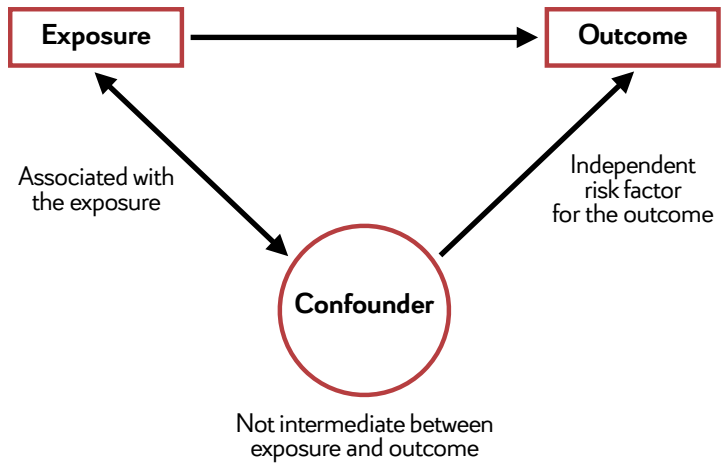


Figure 201. Confounder model.

Conditions Necessary for Confounding

There are **three conditions** that must be present for a variable to behave as a confounder.

1. The confounding factor must be **associated with both** the risk factor of interest and the outcome.
2. The confounding factor must be **distributed unequally** among the groups being compared.
3. A confounder **cannot be an intermediary step in the causal pathway** from the exposure of interest to the outcome of interest.

Let's understand these conditions by analyzing the following example:

It is known that modest alcohol consumption is associated with a decreased risk of coronary heart disease, and it is believed that one of the mechanisms by which alcohol causes a reduced risk is by raising HDL levels, the so-called "good cholesterol." Higher levels of HDL are known to be associated with a reduced risk of heart disease. Consequently, it is believed that modest alcohol consumption raises HDL levels, and this in turn, reduces coronary heart disease. In a situation like this, HDL levels are not a confounder of the association between alcohol and heart disease, because it is part of the mechanism by which alcohol produces this beneficial effect. If increased HDL is a consequence of alcohol consumption and part of the mechanism by which it lowers the risk of heart disease, then it is not a confounder.

Another example to understand confounders can be:

If we examine the impact of smoking on the incidence of lung cancer, a variable such as **exposure to asbestos dust** confounds the relation between smoking and cancer. Exposure to asbestos dust and smoking are associated, i.e. there are proportionally more persons exposed to asbestos in the smoking group than in the non-smoking group. In addition, inhaling asbestos dust is a strong cause of cancer of the pleura. Cancer is the outcome variable in this example, smoking a potential cause, and exposure to asbestos a confounder.

Variables that Can Be Confounders

Not surprisingly, since most diseases have multiple contributing causes (risk factors), there are **many possible confounders**.

- » A confounder can be **another risk factor for the disease**.
 - For example, in a hypothetical cohort study testing the association between exercise and heart disease, age is a confounder because it is a risk factor for heart disease.
- » A confounder can also be a **preventive factor for the disease**.
 - If people who exercised regularly were more likely to take aspirin, and aspirin reduces the risk of heart disease, then aspirin use would be a confounding factor that would tend to exaggerate the benefit of exercise.
- » A confounder can also be a **surrogate or a marker for some other cause of disease**.
 - For example, socioeconomic status may be a confounder in the example of alcohol consumption and risk of coronary disease because lower socioeconomic status is a marker for a complex set of poorly understood factors that seem to carry a higher risk of heart disease.

Controlling Confounders

There are several ways to modify a study design to actively exclude or control confounding variables. The most representative ones include:

- » **Randomization:** The random assignment of study subjects into groups breaks any links between exposure and confounders.
 - This reduces potential for confounding by generating groups that are fairly comparable with respect to known and unknown confounding variables.

» **Restriction:** Eliminates variation in the confounder by selecting only subjects with the same predisposed characteristics (the same age or same sex), eliminating confounders by those characteristics.

» **Matching:** Selection of a comparison group with respect to the distribution of one or more potential confounders (e.g., if age and sex are the matching variables, then a 45 year old male case is matched to a male control with same age).

Eliminating Confounders

To control for confounding in the statistical analysis, investigators should **measure the confounders in the study**. Researchers usually do this by **collecting data on all known, previously identified confounders**.

Once in the analysis stage of the study, there are two options to deal with confounders:

» **Stratification:** Fix the level of the confounders and creates groups within which the confounder does not vary. Then evaluate the exposure-outcome association within each stratum of the confounder.

» **Multivariate models:** Handle large numbers confounders simultaneously.

– This can be achieved with a Logistic Regression, a Linear Regression, or an Analysis of Covariance (ANCOVA).

Simpson's Paradox

The Simpson's Paradox may arise if there is (at least) **one confounding variable that has not been accounted for**. The best way to understand this phenomenon is applying it into an example.

Consider the following scenario to understand the Simpson's paradox:

Suppose the 1st grade students of two medical schools named Alpha and Beta participated in a national standard Neuroanatomy test. We want to compare the average scores of both schools. Assume we are told that the average scores of both male and female in Beta school are higher than those in Alpha school.

What can we say about the overall average score in those schools? Is it true that the Beta School gets a higher average score than the Alpha School?

Table 20.1. Average Scores of Male and Female Students in Alpha and Beta Schools

School	Gender			
	Male		Female	
	n	Average	n	Average
Alpha	80	84	20	80
Beta	20	85	80	81

The answer seems to be affirmative and intuitive. To be more specific, the average scores of male and female students in each school are presented in **Table 20.1**. It is obvious that both male and female students in Beta School have higher average scores. However, simple calculation shows that the overall average scores in both schools are 83.2 and 81.8, respectively. Alpha School won on the average score! Why is this example so counterintuitive? Is there anything wrong here? Is the average score a reasonable measure of the performance of students in a school?

Generally speaking, the Simpson's paradox states that **the conditional relation does not imply marginal relation, and vice versa**. The effect of Simpson's paradox has been way beyond the statistical community. In fact, the Simpson's paradox is very prevalent in many areas, from natural science, to social sciences, and even philosophy.

It can be concluded that it is an **inherent property of data from observational studies**.

Bridge to Observational Studies

In an **observational study**, individuals are observed or certain outcomes are measured. No attempt is made to affect the outcome.

Key Terms

Define the following terms.

Confounder

Confounder control

Confounding

Matching

Multivariate models

Randomization

Restriction

Simpson's paradox

Stratification

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Get together with some classmates and discuss the following:

- Does birth order affect the risk of Down's syndrome?
- Identify the possible confounders.

2. If age was a suspected confounder of the relationship between exercise and pulse, how might you adjust for this factor in your analysis?

3. List the conditions that a variable must meet in order to recognize it as a confounder.

4. List the most representative methods to control confounding variables.

5. Multiple choice questions.

1. A confounder is:

- a) A factor associated with the outcome, but not associated with the exposure.
- b) A factor associated with the outcome and associated with the exposure and on the causal pathway between exposure and outcome.
- c) A factor not associated with the outcome, but associated with the exposure.
- d) A factor associated with the outcome and associated with the exposure, but not on the causal pathway.

2. The NMMAPS was a systematic investigation of the dependence of daily hospitalization and mortality counts on ambient particulate matter (PM) and other air pollutants. The NMMAPS database includes mortality, weather and air pollution data for the largest 90 cities in the USA for 1987 through 2000. Morbidity data was also available for 14 cities that had daily PM10 measurements. Daily data on mortality, weather, and air pollution were obtained from publicly available data sources.

In NMMAPS, a confounding factor is:

- a) Smoking.
- b) Weather patterns.
- c) A factor that varies on short time scale and is associated with daily mortality.
- d) Influenza epidemics.

3. Simpson's Paradox occurs when:

- a) No baseline risk is given, so it is not known whether or not a high relative risk has practical importance.
- b) A confounding variable rather than the explanatory variable is responsible for a change in the response variable.
- c) The direction of the relationship between two variables changes when the categories of a confounding variable are taken into account.
- d) The results of a test are statistically significant but are really due to chance.

4. A study done by the Center for Academic Integrity at Rutgers University surveyed 2116 students at 21 colleges and universities. Some of the schools had an "honor code" and others did not. Of the students at schools with an honor code, 7% reported having plagiarized a paper via the Internet, while at schools with no honor code, 13% did so. Although the data provided are not sufficient to carry out a chi-square test of the relationship between whether or not a school has an honor code and whether or not a student would plagiarize a paper via the Internet, suppose such a test were to show a statistically significant relationship on the basis of this study. What would be the correct conclusion?

- a) Because this is an observational study, it can be concluded that implementing an honor code at a college or university will reduce the risk of plagiarism.

- b) Because this is a randomized experiment, it can be concluded that implementing an honor code at a college or university will reduce the risk of plagiarism.
- c) Because this is an observational study and confounding variables are likely, it cannot be concluded that implementing an honor code at a college or university will reduce the risk of plagiarism.
- d) Because this is a randomized experiment and confounding variables are likely, it cannot be concluded that implementing an honor code at a college or university will reduce the risk of plagiarism.
- 5. Why are confounding variables such a problem in research?**
- a) They are difficult for participants to give responses to.
- b) They make questionnaires too long for participants to complete.
- c) They lead to high attrition rates for studies.
- d) They make it difficult to draw conclusions about the relationships between the main variables in the study.

Bibliography and Suggested Reading

- Ananth CV, Schisterman EF. Confounding, Causality and Confusion: The Role of Intermediate Variables in Interpreting Observational Studies in Obstetrics. *American Journal of Obstetrics and Gynecology*. 2017;217(2):
- Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. *Psychosom Med*. 2004; 66:868-75.
- McDonald JH. *Handbook of Biological Statistics*. 3rd Edition. Sparky House Publishing, Baltimore, Maryland; 2014.
- Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*. 2012;5(2):79-83.
- Parfrey PS, Barret BJ. *Clinical Epidemiology. Practice and Methods*. 2nd Edition. New York: Springer; 2015.
- VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019; 34(3): 211–219.
- Wang B, Wu P, Kwan B, Tu XM, Feng C. Simpson's Paradox: Examples. *Shanghai Arch Psychiatry*. 2018; 30(2): 139-143.
- Wunsch G. Confounding and control. *Demographic Research*. 2007;16(4):97-120.

Attributable Risk

Learning objectives for this chapter

- A. Define the concept of attributable risk.
- B. Calculate and interpret the attributable risk in an exposed group.
- C. Calculate and interpret the population attributable risk.
- D. Describe how attributable risk is used to estimate the potential for intervention.
- E. Understand and interpret the concept of number needed to treat (NNT) and how it is calculated.

Attributable Risk (AR) is a measure of the **proportion** of the disease occurrence that can be **attributed to a certain exposure**. Furthermore, it can correspond to different things:

- » The portion of disease rate that is attributable to the exposure factor in the **epidemiological context**.
- » The portion of correct diagnosis rate attributable to a positive predictive result (e.g. lab test) in the **clinical context**.
- » The portion of beneficial outcome rate attributable to a treatment.

One might determine the proportion of all cases of an outcome in the total population that could be attributed to the exposure to a risk factor. This is called the **Population Attributable Risk (PAR)**. When expressed as a percent, it is called the **Population Attributable Risk Percent**.

Both AR and PAR are mathematical or algebraic assessments of **statistical association**, but they **do not provide any information on causal relationship** (see **Appendix B. Bradford Hill's Criteria for Causality**).

Attributable risk helps you determine how much of an outcome may be attributable to a particular risk factor in a population exposed to that factor.

Calculating the population attributable risk percent allows you to determine what percent of an outcome could possibly be prevented if a risk factor were to be removed from the population.

How to Calculate an Attributable Risk

Remember that, when conducting a risk analysis, epidemiologists begin by constructing a 2x2 table as illustrated in **Table 19.2**, where “a” represents those in the exposed population who experienced the outcome of interest and “b” those in the exposed population who did not experience that outcome. In this case, the **risk of exposure** is expressed as “ $a/a+b$ ”. Conversely the **risk for those not exposed** to the risk factor would be “ $c/c+d$ ”.

To calculate the AR, you simply subtract the risk for the non-exposed group (p_2) from the risk for the exposed group (p_1) with the following **formula**:

$$\text{Attributable risk (AR)} = p_1 - p_2$$

How to Calculate a Population Attributable Risk

PAR depends not only on the excess risk imposed by the exposure, but also on the share of the total population that is exposed. **Two formulas** can be used to calculate the PAR:

$$\text{PAR} = p_0 - p_2$$

Or:

$$\text{PAR} = (p_1 - p_2) \times n_1/N$$

Where:

- » p_0 = the proportion of ALL cases with the outcome of interest ($a + c/a + b + c + d$).
- » p_1 = the proportion of cases with the outcome of interest WITH the exposure to the risk factor ($a / a + b$).
- » p_2 = the proportion of cases with the outcome of interest WITHOUT the exposure to the risk factor ($c / c + d$).
- » N = the total number of cases ($a + b + c + d$).
- » n_1 = the number of cases exposed to the risk factor ($a + b$).

PAR is a measure of the **magnitude** of a given problem from a **Public Health point of view**, e.g., the proportion of lung cancers that can be attributed to smoking:

- » If the prevalence of smoking is set to 50%, or 0.5 of the population, and the relative risk of lung cancer is set at 10, this will give us a PAR of 0.82.

How to Interpret an AR?

Based on the previous example on lung cancer and smoking, an AR of **0.82** implies that **82% of all lung cancer in the population can be attributed to smoking**.

AR can be used to determine the **potential impact of prevention** or health promotion if the prevalence of the exposure is reduced.

Another example

Table 21.1 summarizes the results of a study population by Iso H, et al. of 41,782 men aged 40–79 years living in 45 communities across Japan from 1988 to 1990 and followed through the end of 1999.

The incidence rate per 100,000 person-years of cardiovascular disease among current smokers is 399 and among non-current smokers is 356; overall the rate is 379. The rate ratio (risk ratio) is 1.122, meaning that male current smokers are 1.122 times (or 12.2%) more likely than nonsmokers to develop cardiovascular disease.

- » **Attributable risk:** Calculated as the difference in attack rates ($I_e - I_o$).
 - In this case, the AR is 43.6 per 100,000. That is, the excess occurrence of cardiovascular disease among male smokers that can be attributed to their smoking is 43.6 per 100,000.
- » **Attributable risk percent:** Calculated with I_e and I_o .
 - In this case, the AR percent is 10.9% $[(1.122 - 1)/1.122 \times 100]$. That is, if smoking causes cardiovascular disease, nearly 10.9% of cardiovascular disease in males who currently smoke is attributable to their smoking.
- » **Population attributable risk:** Calculated by subtracting the person-time rate in the unexposed group from the person-time rate in the total population.
 - In this case, if current smoking were eliminated from the population, the cardiovascular disease incidence rate could be expected to drop by 23 per 100,000.

Table 21.1. Total Cardiovascular Disease Based on Smoking Status

	Cases	Controls	Person-Years
Current smoker	882	–	220,965
Nonsmoker	673	–	189,254
Total	1,555	–	410,219

» **Population attributable risk percent:** In this example, it is 6.2% $[(379 - 356)/379 \times 100$.

– This means that, if smoking were eliminated from the population, a 6.2% decrease in the incidence rate of cardiovascular disease could be expected.

The Number Needed to Treat (NNT) is the number of patients you need to treat to prevent one additional bad outcome (death, stroke, etc.).

Number Needed to Treat (NNT)

The NNT concept allows us to attach concrete numbers to concepts of relative risk (RR) and AR (or risk difference). NNT is an **absolute effect measure** which is interpreted as **the number of patients needed to be treated with one therapy versus another for one patient to encounter an additional outcome of interest within a defined period of time**.

The calculation of NNT is founded on the cumulative incidence of the outcome per number of patients followed over a given period of time, being classically calculated by inverting absolute risk reduction (ARR) (also called risk difference [RD]) between two treatment options. It can be done with the following **formula**:

$$\text{NNT} = 1/\text{ARR}$$

The resulting value of the NNT is **specific to a single comparison** between two treatment options within a single study, rather than an isolated absolute measure of clinical effect of a single intervention.

The calculation of NNT should be based upon the use of methods that **align with the characteristics of a given study**, such as the research design and the type of variable (e.g., binary, time to event, or continuous) used to express the outcome of interest.

How to Interpret a NNT?

Once again, it will be easier to understand this concept if we use an example.

Suppose that the RR might be 2 for each of two different diseases.

» For disease A, the AR might be 0.1, so that NNT would be $1/0.1 = 10$.

» For disease B, the AR might be 0.001, so that NNT would be $1/0.001 = 1000$.

For disease A, we would need to treat 10 patients for one to benefit. On the other hand, for disease B we would need to treat 1000 patients for one to benefit.

Key Terms

Define the following terms.

Absolute risk reduction
Attributable risk

Attributable risk percent
Number needed to treat

Population attributable risk
Population attributable risk percent

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Get together with some classmates and discuss the following:

- Is population attributable risk relevant in a public health context?

2. The serious potential burden of iatrogenic disease caused by hormone replacement therapy (HT) use is shown by research indicating that the population attributable risk of breast cancer due to HT likely ranges between 10 to at least 20 per cent, which translates to an excess burden of breast cancer cases in the past decade.

- Discuss with some classmates how ignoring health inequities can lead to invalid epidemiological findings and harm the public's health.

3. State the differences between relative and attributable risks.

4. True or false: Attributable risk provide information to infer causality.

5. In a wound infection study, the incidence in the exposed group was 5.3 per 100. Of this, 4 per 100 could be attributed to having had the incidental appendectomy (the other 1.3 per 100 was the "inherent risk" of the staging laparotomy).

- What percentage of the wound infections in the group that had the incidental appendectomy could be attributed to having

had the appendectomy?

6. In 1995, Moore conducted a study to assess the effectiveness of the triple therapy in Helicobacter pylori treatment, in comparison to Histamine antagonist. Two outcomes were proposed: (1) eradication of Helicobacter pylori after 6-10 weeks of treatment; (2) ulcers remaining cured at 1 year after 6-10 weeks of treatment. Two NNTs were identified: 1.1 and 1.8, corresponding respectively to outcome (1) and (2).

- Discuss with some classmates: How would the NNT would change if the outcome (2) is not ulcers-remaining-cured-at-1-year but ulcers-remaining-cured-at-2-years?

7. A total of 360 patients participated in a randomized controlled trial designed to compare the effectiveness of drug X in reducing deaths with a placebo. Out of 120 patients in the treatment group, 12 patients died within three years. Out of 240 patients in the control group, 48 patients died within three years.

- What is the number needed to treat to prevent 1 death?

Bibliography and Suggested Reading

- Bewick V, Cheek L, Ball J. Statistics review 11: Assessing risk. *Critical Care*. 2004;8(4):287-291.
- Hazra A, Gogtay N. Biostatistics series module 8: Assessing risk. *Indian J Dermatol*. 2017 Mar-Apr;62(2):123-129.
- Hoffman JIE. Odds Ratio, Relative Risk, Attributable Risk, and Number Needed to Treat. In: Hoffman JIE. *Basic Biostatistics for Medical and Biomedical Practitioners*. Waltham: Elsevier Academic Press; 2019.
- Iso H, Date C, Yamamoto A, et al. Smoking cessation and mortality from cardiovascular disease among Japanese men and women: the JACC study. *Am J Epidemiol*. 2005;161(2):170–179.
- Mendes D, Alves C, Batel–Marques F. Number needed to treat (NNT) in clinical literature: an appraisal. *BMC Medicine* (2017) 15:112.
- Richard T, Vanhaeverbeek M; Meerhaeghe AV. The Number Needed to Treat (NNT). *Rev Med Brux*. Sep-Oct 2011;32(5):453-8.
- Riffenburgh RH. Tests on Categorical Data. In: Riffenburgh RH. *Statistics in Medicine*. 3rd Edition. Burlington: Elsevier Elsevier Academic Press; 2012.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. *Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication*. Waltham: Elsevier Academic Press; 2015.

Section V

Use of Diagnostic Tests

Chapters of the Section

Chapter 22	Likelihood Ratios
Chapter 23	Pre and Post-test Probability
Chapter 24	Sensitivity, Specificity, and Predictive Values
Chapter 25	ROC Curves

Likelihood Ratios

Learning objectives for this chapter

- A. Define likelihood ratio (LR).
- B. Calculate both the positive and negative LR for a diagnostic test.
- C. Interpret what the LR represents in a clinical scenario.
- D. Learn about the Fagan Nomogram as a tool to determine a patient's probability of having disease based on a test result.
- E. Understand the concept of diagnostic thresholds and apply it to clinical practice.

Despite their usefulness in interpretation of clinical findings, laboratory tests, and imaging studies, **Likelihood Ratios** (LR) are less used. Based on that, we will try to make LR both simple and clinically relevant, trying to enhance your familiarity with and use of LR.

LR constitute one of the best ways to **measure and express diagnostic accuracy**. LR is the **likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that the same result would be expected in a patient without the target disorder**.

LR can be obtained by **dividing the probability** of a finding in patients with disease divided by the probability of the same finding in patients without disease.

» For example, among patients with abdominal distension who undergo ultrasonography, the physical sign “bulging flanks” is present in 80% of patients with confirmed ascites and in 40% without ascites (i.e., the distension is from fat or gas). The LR for “bulging flanks” in detecting ascites, therefore, is 2.0.

The **implications** are clear: ill people should be much more likely to have an abnormal test result than healthy individuals.

Furthermore, LR measures the **power** of a test to change the **pre-test** into the **post-test probability** of a disease being present (more on this in the following chapter).

As its name implies, LR is the likelihood of a given test result in a person with a disease compared with the likelihood of this result in a person without the disease.

Another way to calculate the LR is:
Sensitivity/(1-specificity)

LRs can also be calculated from a **2 x 2 table**, as shown in **Figure 22.1**. Let's learn how to do so by analyzing the following example:

Suppose you are in charge of 15 people who are sick. 12 of these (80%) have a true-positive test for the disease. On the other hand, you also have 85 patients who are healthy, but 5 of those (6%) have a false-positive test.

» The LR for a positive test is simply the ratio of both percentages (80%/6%), which is 13.

– Stated in another way, people with the disease are **13 times more likely** to have a positive test than are those who are healthy.

– This is called the **positive likelihood ratio** (abbreviated LR+)

» The LR for a negative test (called **negative likelihood ratio** or LR-) is calculated similarly. Three of 15 sick people (20%) have a false-negative test, whereas 80 of 85 healthy individuals (94%) have a true-negative test.

– So LR- is the ratio of these percentages (20%/94%), which is 0.2. Thus, a negative test is a fifth **as likely** in someone who is sick than in a well person.

		Disease		
		Present	Absent	
Test	Positive	a	b	a + b
	Negative	c	d	c + d
		a + c	b + d	
LR+ = 0.80/0.06 = 13		12 (80%)	5 (6%)	17
LR- = 0.20/0.94 = 0.2		3 (20%)	80 (94%)	83
		15 (100 %)	85 (100%)	

Figure 22.1. 2x2 table used to calculate LR (top), 2x2 table with the example of how to calculate LR (bottom).

As you could infer from the last example, the **LR of a positive test** tells us how well a positive test result does by comparing its performance when the disease is present with when the disease is absent.

On the other hand, the **LR of a negative test** tells us how well a negative test result does by comparing its performance when the disease is absent with when the disease is present.

Fagan Nomogram

The **Fagan nomogram** (Figure 22.2) is a convenient tool that shows how a test that has a known LR can **change the pre-test probability**.

» In this nomogram, a straight line drawn from a patient's pre-test probability of disease (which is estimated from experience, local data or published literature) through the LR for the test result that may be used, will point to the post-test probability of disease.

The best test to use for ruling in (accepting) a disease is the one with the largest likelihood ratio of a positive test.

The best test to rule out (discard) a disease is the one with the smaller likelihood ratio of a negative test.

The Fagan nomogram is rarely accessible at the bedside, and is seldom used. But keep it in mind just in case you need this information in an upcoming evaluation.

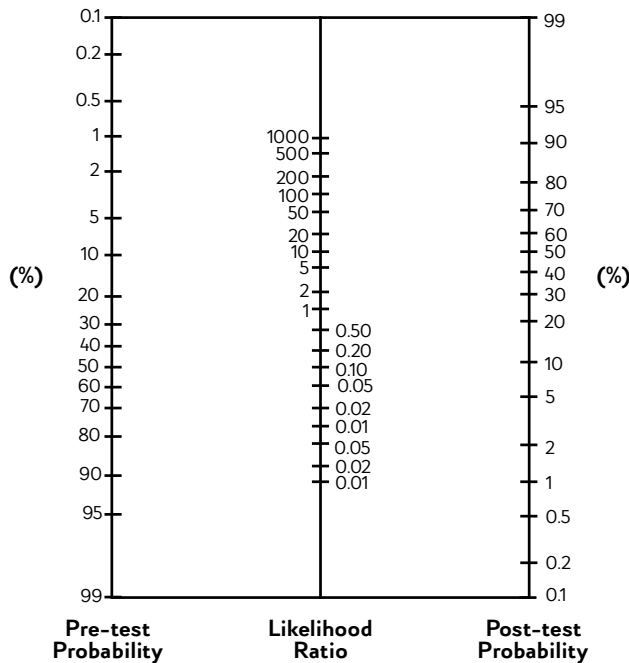


Figure 22.2. The Fagan nomogram.

Sensitivity, and Specificity



Sensitivity: the ability of a screening test to detect a true positive, reflecting a test's ability to correctly identify all people who have a condition.

Specificity: the ability of a screening test to detect a true negative, reflecting a test's ability to correctly identify all people who do not have a condition.

Interpretation of LR

The LR is a way to incorporate the **sensitivity** and **specificity** of a test into a single measure, but is independent of the prevalence of the disease in the population.

LR is a **ratio of likelihoods** (or probabilities) for a given test. The first is the probability that a given test result occurs among people with disease. The second is the probability that the same test result occurs among people without disease. The ratio of both probabilities (or likelihoods) is the LR.

- » A **LR greater than 1** increases the probability that the disease of interest is present, and the higher the LR, the greater increase in probability.
- » A **LR less than 1** decreases the probability of the target disorder, and the smaller the LR, the greater the decrease in probability.

How Much do LRs Change Disease Likelihood?

LRs of different sizes have different **clinical implications**: The further the LR is from 1.0, the greater its effect on the probability of disease. This can be summarized in **Table 22.1**.

For example, a LR greater than 10 or less than 0.1 generates large and often conclusive changes from pre-test to post-test probability. On the other hand, LRs of 1 to 2 and 0.5 to 1 alter probability to a small (and rarely important) degree.

Calculating LR for Tests With Two Outcomes

The simple 2x2 tables in **Figure 22.1** shows how to calculate LR for test with two outcomes.

Table 22.1. How Much do LRs Change Disease Likelihood?

LRs greater than 10 or less than 0.1	Cause large changes
LRs 5–10 or 0.1–0.2	Cause moderate changes
LRs 2–5 or 0.2–0.5	Cause small changes
LRs less than 2 or greater than 0.5	Cause tiny little changes
LRs = 1.0	Cause no change at all

Calculating LR for Tests With Multiple Outcomes

The calculation of this kind of LR is similar to the calculation for dichotomous outcomes.

In this case, a separate LR is calculated **for every level of test result**. Again, let's better understand this with an example:

Table 22.2 summarizes white-blood-cell counts for 59 patients with appendicitis and 145 patients without the diagnosis. To calculate the LR for a count of 7×10^9 cells per L, 2% is the numerator (those with appendicitis) and 21% the denominator (those without appendicitis); the likelihood ratio is $2\%/21\%$, or 0.1. This same calculation is done for every level of white-blood-cell count; for the highest values, the calculation cannot be done because the denominator is zero. Likelihood ratios vary from 0.1 to infinity, with a trend towards higher ratios with higher white-blood-cell counts.

An Useful Mnemonic

Regrettably, nomograms and computers are usually not available at the bedside. Hence, a mnemonic suggested by **McGee** for simplifying the use of LRs has strong appeal. He notes that:

» For pre-test probabilities between 10–90% (the usual situation), the change in probability from a test or clinical finding is approximated by a **constant**.

Table 22.2. Likelihood Ratios for White-blood-cell Count in Diagnosing Appendicitis

Cells per L	n (%) with appendicitis	n (%) without appendicitis	% with appendicitis/ %without appendicitis	LR
$\leq 7 \times 10^9$	1(2%)	30(21%)	2/21	0.10
7–9 $\times 10^9$	9(15%)	42(29%)	15/29	0.52
9–11 $\times 10^9$	4(7%)	35(24%)	7/24	0.29
11–13 $\times 10^9$	22(37%)	19(13%)	37/13	2.8
13–15 $\times 10^9$	6(10%)	9(6%)	10/6	1.7
15–17 $\times 10^9$	8(14%)	7(5%)	14/5	2.8
17–19 $\times 10^9$	4(7%)	3(2%)	7/2	3.5
$\geq 19 \times 10^9$	5(8%)	0	8/0	∞
Total	59(100%)	145(100%)		

Table 22.3. Likelihood Ratios and Bedside Estimates

LRs between 0 and 1 reduce the probability of disease	Approximate change in probability (%)
0.1	-45
0.2	-30
0.3	-25
0.4	-20
0.5	-15
1.0	0
LRs greater than 1 increase the probability of disease	Approximate change in probability (%)
2	+15
3	+20
4	+25
5	+30
6	+35
7	
8	+40
9	
10	+45

» As clinician, you need to remember only three benchmark likelihood ratios: **2, 5, and 10** (Table 22.3), which correspond to the first three multiples of 15%:

- A LR of 2 increases the probability by about 15%.
- A LR of 5 increases the probability by about 30%.
- A LR of 10 increases the probability by about 45%.

For example, with a pre-test probability of 40% and a LR of 2, the post-test probability is $40\% + 15\% = 55\%$ (quite close to the 57% when calculated by formula).

For **LR less than 1**, the rule works in the opposite direction.

- » The reciprocal of 2 is 0.5; that of 5 is 0.2, and that of 10 is 0.1.
 - A LR of 0.5 would reduce the pre-test probability by about 15%.
 - A LR of 0.1 would reduce the pre-test probability by about 45%.

Diagnostic Thresholds

Diagnostic tests should only be used when they **will modify our planned management**. If a clinician's pre-test probability of a disease securely rules in or out a diagnosis, further testing is **unwarranted**. More testing should be considered only in the **murky middle zone of clinical uncertainty** (Figure 22.3).

The location of the decision thresholds (A and B) along the continuum of diagnostic certainty needs to be determined **before testing is done** and must be tailored to the specific patient in question.

Based on Figure 22.3, probabilities lower than point A effectively **exclude the diagnosis** in question.

» Hence, point A becomes the **testing threshold**: Pre-test probabilities greater than A but lower than B could benefit from further testing.

Probabilities higher than point B **justify beginning treatment without further delay**.

» Hence, point B becomes the **treatment threshold**.

Using LR in Daily practice

The main advantage of LRs (over other measures of diagnostic accuracy, such as sensitivity and specificity) is that clinicians can use them to **quickly compare different diagnostic strategies** and thus **refine a clinical judgment**. Several types of these comparisons can be analyzed in Table 22.4.

The most common comparison in which LRs are used is in examining **different tests for the same diagnosis**.

LRs also convey the point that positive and negative results of the same test often change probability asymmetrically.

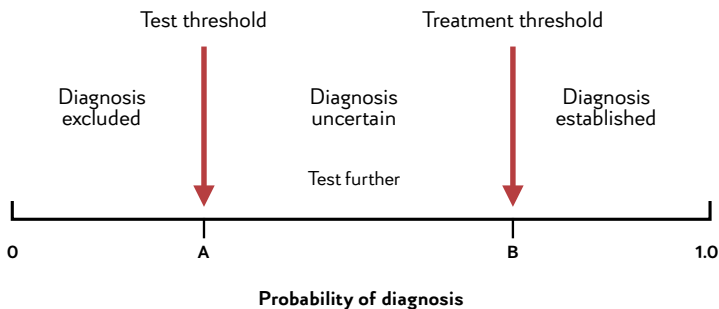


Figure 22.3. Thresholds for testing and treating, as a function of probability of diagnosis.

Table 22.4. Types of Comparisons Using LR_s

Finding	LR ₊	LR ₋
Compare the accuracy of different diagnostic tests for the same diagnosis (physical findings detecting ascities)		
Bulging flanks	2.0	0.3
Flank dullness	2.0	0.3
Edema	3.8	0.2
Fluid wave	6.0	0.4
Shifting dullness	2.7	0.3
Compare the diagnostic impact of a test's positive result to its negative result		
Hyperresonance, detecting chronic airflow obstruction	51	NS
Pleuritic component to chest pain detecting myocardial infarction	0.2	1.3
Compare the accuracy of the same test for different definitions of disease (late-peaking systolic murmur)		
Detecting severe aortic stenosis	3.0	0.2
Detecting moderate-to-severe aortic stenosis	24.2	0.3
Compare the accuracy of different levels of the same test for the same diagnosis (CAGE questions for detecting alcohol abuse and dependence)		
0 positive	0.1	–
1 positive	NS	–
2 positive	4.5	–
3 positive	13.3	–
4 positive	101	–
1 or more points	4.7	0.1
Compare the accuracy of the same test for the same diagnosis in different clinical settings (tachypnea detecting pneumonia in children)		
All children	2.2	0.4
Disease duration <3 days	NS	NS
Disease duration 3–5 days	NS	NS
Disease duration ≥6 days	3.5	0.1

*NS = not significant

Clinicians may also use LRs to compare the accuracy of the **same test for different definitions of disease**, which usually provides important insights into the value of the test and its limitations. Furthermore, LRs may also compare **different levels of findings for the same diagnosis**. This type of comparison is possible if the finding can be measured and placed in 3 or more levels (i.e., the finding is not simply “present” or “absent”).

Finally, LRs may examine the diagnostic accuracy of the same test for the same disease but when **applied to different clinical settings**, a comparison that identifies in which **group of patients a test is most discriminatory**.

Key Terms

Define the following terms.

Diagnostic accuracy	Likelihood	Positive likelihood ratio
Diagnostic threshold	Likelihood ratio	Test threshold
Fagan nomogram	Negative likelihood ratio	Treatment threshold

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- The rapid antigen detection test for streptococcal pharyngitis has a positive and negative likelihood ratio of 35 and 0.3, respectively. You are called to the urgent care center to evaluate a child who has sudden onset of sore throat, exudative pharyngitis, and a fever of 103 °F. You estimate the pretest probability of streptococcal pharyngitis in this child to be 60%.

 - If the rapid antigen test is positive, what is the probability that this child has actual streptococcal pharyngitis? Use the Fagan nomogram.
- Interpret the following LR+ and LR–:

 - Homan’s sign: LR+ 1.5; LR– 0.6.
 - Palmar pallor in anemia: LR+ 5.6; LR– 0.4.
 - Phalen sign: LR+ 1.3; LR– 0.7.
 - Babinski sign for unilateral cerebral hemisphere disease: LR+ 8.5; LR– not significant.
 - Rectal temperature for detecting infection: LR+ 6.1; LR– 0.6.
 - Unability to perform 10 tandem steps in diagnosing Parkinson’s Disease: LR+0.2; LR– 5.4.
 - Capillary refill time ≥ 5 s for predicting multi-organ dysfunction: LR+2.6; LR– 0.3.
 - Heart rate >95 beats/min for predicting hospital mortality in septic shock: LR+2.0; LR+0.1.

3. Complete the **Table AL22.1** with the interpretation of the different likelihood ratios.

Table AL22.1. Interpreting Positive and Negative Likelihood Ratios

Positive Likelihood Ratio (LR+)			Negative Likelihood Ratio (LR-)		
Value	Descriptor	Interpretation	Value	Descriptor	Interpretation
≥10	Very positive		≤0.10	Very negative	
3	Moderately positive		≤0.30	Moderately negative	
1	Neutral		1	Neutral	

Bibliography and Suggested Reading

- Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatr.* 2007 Apr;96(4):487-91.
- Attia J. Moving beyond sensitivity and specificity: using likelihood ratios to interpret diagnostic tests. *Aust Prescr.* 2003; 26: 111-113.
- Crewe S, Rowe C. Likelihood Ratio in Diagnosis. *Pediatrics in Review.* 2011;32(7):296-298.
- Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med.* 1975; 293(5):257.
- Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet.* 2005; 365: 1500-05.
- McGee S. Diagnostic Accuracy of Physical Findings. In: McGee S. *Evidence-Based Physical Diagnosis.* Philadelphia: ELSEVIER; 2012.
- McGee S. Simplifying Likelihood Ratios. *J Gen Intern Med.* 2002; 17(8): 647-650.
- Richardson WS, Wilson MC, Keitz SA, Wyr PC. Tips for Teachers of Evidence-based Medicine: Making Sense of Diagnostic Test Results Using Likelihood Ratios. *J Gen Intern Med.* 2008 Jan; 23(1): 87-92.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: How to practice and teach EBM.* 2^o Edition. Edinburgh: Churchill Livingstone; 2000.

Pre and Post-test Probability

Learning objectives for this chapter

- A. Use and define clinical prediction rules.
- B. Define pre and post-test probability.
- C. Practice applying pre and post-test probability in clinical decision-making scenarios.
- D. Demonstrate how the estimate of pre-test probability influences the interpretation of diagnostic tests and patient management.

Before we dive into this topic, consider the following scenario:

A 3-day-old patient is referred to your hospital, and you are on call. Assume that you know nothing more than the age. The list of possible diagnoses is enormous: congenital anomalies of the bowel, urinary tract infections, seizures, aspiration pneumonia, heart disease, and so on. You examine the patient:

Suppose that the patient is cyanotic. Then, heart and lung diseases go to the top of the list of possible diagnoses.

If the baby is breathing normally, then heart disease is more likely than lung disease.

If there is no cardiac murmur, then transposition of the great arteries goes to the top of the list.

And so on...

Every clinical encounter begins with an **initial clinical impression**. This is a **subjective probability** of disease. The ultimate goal of performing a diagnostic test is to **refine** this probability to the point where the physician can confidently make a **treat or no-treat decision**.

Thus, pre and post-test probabilities provide a **framework** that helps us achieve this decision, whether we treat, test for, or toss specific diagnoses, helping us to prioritize our differential.

Probability in Clinical Scenarios

Understanding how specific **tests** alter the relative probabilities of our diagnostic impression is an important aspect of clinical reasoning. In order to do that, we need to build a **mental map** of how likely different diseases are, and how those probabilities shift as we gather and synthesize data.

As a physician, one of the crucial things to perform once you get a new case, is to adequately **connect book-knowledge to that particular patient**. By combining both things, we are performing a clinical reasoning in our heads in order to come up with a diagnosis and a treatment plan.

When you match all the knowledge stored in your hippocampus about a specific disease with the patient in front of you, a series of **differential diagnosis** comes up. Each of this diagnoses gets assigned a specific probability that falls somewhere on a **spectrum** between 0% and 100%. In Medicine it is very difficult to determine a 0% and a 100% probability. That's why probability may be classified into **3 different areas**:

- » **Treatment area:** probability $\geq 80\%$.
- » **Test area:** $\geq 20-79\%$
- » **Trash area:** $\leq 19\%$.

Furthermore, Catherine R Lucey, M.D. suggested that this spectrum of probability may be subdivided into **five categories** based on the suspicion that a patient has a disease, as shown in **Table 23.1**.

Catherine Reinis Lucey, M.D., is Vice Dean for Education and Executive Vice Dean for the School of Medicine at the University of California, San Francisco (UCSF). She directs the undergraduate, graduate and continuing medical education programs of the School of Medicine and the Office of Medical Education.

Table 23.1. Spectrum of Probabilities by Catherine R Lucey, M.D.

"Pathognomonic" feature	Very likely	90–100%
Discriminating feature	Likely	80–90%
		67–80%
	Uncertain	60–66%
		50–60%
		40–50%
		33–40%
Has a few differentiating features	Unlikely	20–32%
		10–20%
Has rejecting features	Very unlikely	0–10%

Summarizing, the **probability** in each of the diagnoses that you'll come up with will be based on:

- » The patient's clinical presentation.
- » The patient's medical history (risk factors).
- » The base rate of the suspected disease.

Pre and Post-test Probabilities

The **pre-test probability** is the probability you give to each of the potential diagnoses that you've come up with of being the correct diagnosis **before you collect more data**.

- » **Pre-test odds** can be calculated with the following formula: $\text{Pre-test probability} / [1 - \text{pre-test probability}]$.

When new data is obtained, either by performing imaging or laboratory tests, your thinking about the case, as well as the diagnostic likelihood of the disease, will **change**. This new framing is the **post-test probability** or the probability of each diagnosis after initial diagnostic tests have returned.

- » **Post-test odds** can be calculated with the following formula: $\text{Pre-test odds} * \text{Likelihood Ratio}$.
- » **Post-test probability** can be calculated with the following formula: $\text{Post-test odds} / [\text{post-test odds} + 1]$

Diagnostic reasoning is an **iterative process** over **multiple rounds of testing**. That means that the post-test probability obtained from your first aliquot of data will become a new pre-test probability for the next round of testing.

How do Pre-test Probabilities Guide Clinical Decisions?

When considering the probability that a given disease is causing our patient's symptoms, there are essentially **three potential outcomes** for that possible diagnosis:

1. You may think that the probability of this disease is **so LOW** that you can **toss it off** of our differential list.
2. You may think that the probability of this disease is **somewhere in the MIDDLE**, so you need to perform **further testing**.
3. You may think that the probability of this disease is **so HIGH** that you can decide to **treat it**.

It should be noted, however, that the exact **thresholds** for “toss, test, and treat” (the “3 T’s”) change based on factors **beyond probability alone**. Other factors to take into account are:

- » The **morbidity of the disease**: You’re more likely to test for a “can’t miss” diagnosis even if your pre-test probability is very low.
- » The **morbidity of the treatment**: There is a higher threshold of diagnostic certainty to start chemotherapy for cancer than for antibiotics in a possible infection.
- » **Patient preferences**.

As a General Principle

- » The **higher the pre-test probability** for a certain disease, the **better the test** has to be in order to **rule out** the disease.
- » The **lower the pre-test probability** for a certain disease, a **test that**, when positive, **is strongly suggestive** of the specific diagnosis is needed in order to meaningfully increase the pre-test probability.

Post-test probabilities support the physician in a way that you can discriminate which test is best for the patient in terms of **costs** and **safety**, and by which an acceptable post-test probability can be achieved.

Bayes’ Theorem

Recall the 3-day-old patient at the beginning of this chapter. Almost all of the physicians use a **Bayesian approach** to medical diagnosis, a procedure, often so quickly, that we do not realize that we are using a sequential logical process. What Bayes’ theorem does is to **formalize** and **quantify** this process.

The Bayes’ theorem, named after Reverend Thomas Bayes, an 18th century mathematician, describes how to **update or revise beliefs in the light of new evidence**. Applied to diagnostic tests, the theorem describes **how the result of a test** (positive or negative) **changes our knowledge of the probability of disease**.

This is done by **combining** the **pre-test probability** of disease with the **likelihood ratio** (discussed in **Chapter 22**) of the test.

In routine clinical practice, there are two ways of using the Bayes’ theorem to estimate **post-test probability**:

- » By **mathematical calculation**.
 - Obtained by **multiplying** the pre-test odds by the likelihood ratio of the test.
- » By using the **Fagan’s nomogram (Figure 22.2)**.

Special Considerations on the Bayes' Theorem

Bayes' theorem helps overcome many well-known cognitive errors in diagnosis, such as **ignoring the base rate**, **probability adjustment errors** (conservatism, anchoring and adjustment), and **order effects**.

A diagnosis is not necessarily confirmed just because a test was positive. Diagnosis is usually not a binary decision, but a dynamic **probabilistic assessment**, as mentioned before. Bayes' theorem, combines the pre-test probability, the test result (positive or negative) and the sensitivity and specificity (more on this in the next Chapter), or LRs to produce the post-test probability of the disease.

A positive test increases confidence in a diagnosis, but usually **does not indicate certainty**. Whether this confidence exceeds a treatment (or action) threshold remains a decision for the clinician. Likewise, a negative test decreases confidence in a diagnosis, **but rarely rules it out completely**. It is up to those responsible of decision-making to decide if further action is warranted.

Qualitative Procedure to Approximate the Results of a Bayesian Diagnostic Decision Analysis

Bayes' theorem's concepts can be applied using qualitative methods. Here's a procedure guide that can be followed in order to do so.

1. What is the **pre-test probability of the disease** being considered (likely, uncertain or unlikely)?
 - Ideally, this comes from an evidence-based source.
 - If it is very likely (<10–20%) or very unlikely (>80–90%), in general no further testing is needed.
2. If the test is **positive**, the post-test probability increases by one qualitative category (e.g., unlikely to uncertain).
 - If the test is **negative**, the post-test probability decreases by one qualitative category (e.g., unlikely to very unlikely).
3. Perform this process until you are comfortable enough with the **confidence** in the diagnosis, always considering the patient's preferences, the risk of the disease, and the effects of treatment.
4. Negative tests with sensitivities near 99% can almost certainly **rule out** a disease, since the post-test sensitivity will be very unlikely even if the original pre-test probability was likely.
 - Positive tests with specificities near 99% can almost certainly **rule in** a disease.

5. If the pre-test probability was very likely or very unlikely, and further testing is indicated, **two tests** are needed to escape the “very” categories.

– This is because the change in the probabilities is small within these categories.

Key Terms

Define the following terms.

Baye’s theorem

Diagnostic reasoning

Post-test odds

Post-test probability

Pre-test odds

Pre-test probability

Spectrum of probability

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. List other factors that help to determine the “toss, test, and treat” thresholds.

2. Imagine that a research study stated that 10% of people over 50 suffer a particular type of arthritis. A new method to detect the disease was developed and after clinical trials researchers observed that if the method was applied to people with arthritis, they would get a positive result in 85% of cases, while if the method was applied to people without arthritis, they would get a positive result in 4% of cases.

- What is the probability of getting a positive result after applying the method to a random person?
- If the result of applying the method to one person has been positive, what is the probability of having arthritis?

3. If the prevalence of disease X among a certain population is 25%, the pre-test probability of this disease will be 0.25. From this:

- Calculate the pre-test odds.
- If the likelihood ratio of this test is 10, calculate the post-test odds.
- Convert the post-test odds into post-test probability.
- Interpret the results.

4. The likelihood ratio of ultrasonography in detection of traumatic lens dislocation was estimated to be 49.5 in a study by Haghighi et al. Considering 15% prevalence of lens dislocation in an example population:

- Calculate post-test probability of lens dislocation in patients with unilateral blindness following direct eye trauma.

5. Assume that the prevalence of a certain disease is 25% and the positive and negative likelihood ratios of the chosen test are 5 and 0.4, respectively.

- Use the Fagan’s nomogram from **Figure 22.2** to calculate the post-test probability.

Bibliography and Suggested Reading

- Bois FY. Bayesian inference. *Methods Mol Biol.* 2013. 930:597-636.
- Hoffman JIE. Probability, Bayes Theorem, Medical Diagnostic Evaluation, and Screening. In: Hoffman JIE. *Basic Biostatistics for Medical and Biomedical Practitioners.* Waltham: Elsevier Academic Press; 2019.
- Mayor D. *Essential Evidenced-Based Medicine.* United Kingdom: Cambridge University Press; 2004.
- Medow MA, Lucey CR. A qualitative approach to Bayes' theorem. *Evid Based Med.* 2011 Dec;16(6):163-7.
- McGee S. Diagnostic Accuracy of Physical Findings. In: McGee S. *Evidence-Based Physical Diagnosis.* Philadelphia: ELSEVIER; 2012.
- Ojaghi Haghighi SH, Morteza Begi HR, et al. Diagnostic Accuracy of Ultrasound in Detection of Traumatic Lens Dislocation. *Emerg (Tehran).* 2014 Summer; 2(3):121-4.
- Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev.* 2011 Oct 5. CD001431.
- Vos P, Wu Q. Probability Essentials. In: Vinod HD, Rao CR. *Handbook of Statistics 42.* Philadelphia: ELSEVIER; 2018.

Sensitivity, Specificity, and Predictive Values

Learning objectives for this chapter

- A. Define the concepts of sensitivity and specificity in the context of diagnostic medical tests.
- B. Define the concepts of false positive and false negative.
- C. State the relationship of prevalence of disease to the sensitivity, specificity, and predictive values of a diagnostic test.

Sensitivity, specificity, and predictive values are **validity criteria** that quantify the ability of a test to classify a person correctly or erroneously, based on the presence or absence of an exposure or disease. The validity of a measurement is calculated based on the information contained in a **2 x 2 table (Table 24.1)**.

The presence or absence of the disease is determined from a reference criterion which ideally should always be positive in individuals with the disease, and negative in those who do not have it. On the other hand, there is the result of the test, or measurement in general, that you want to evaluate. Most of the times, the measure will be a **dichotomous measure**, so the result will be classified as **positive** or **negative**.

This categorization is established to see whether the outcomes correspond to what is regarded as a **definitive indicator**, often referred to as “**Gold Standard**”, of the same target condition.

Table 24.1. 2 x 2 Contingency Table for the Calculation of Validity Criteria

	Disease	No disease	Total
Test positive	a True positive	b False positive	a + b
Test negative	c False negative	d True negative	c + d
Total	a + c	b + d	a + b + c + d

The term “Gold Standard” suggest that this initial categorization is made on the basis of a test that provides authoritative, **and presumably indisputable evidence**, that a condition does or does not exist.

Sensitivity, Specificity, and Other Terms

The following terms are fundamental to understand the utility of diagnostic tests, and are based on the information in **Table 21.4**.

1. **True positive (a)**: The patient has the disease and the test is positive.
2. **False positive (b)**: The patient does not have the disease, but the test is positive.
3. **True negative (d)**: The patient does not have the disease and the test is negative
4. **False negative (c)**: The patient has the disease, but the test is negative.

Sensitivity

Sensitivity refers to the ability of a test to correctly identify those patients with the disease.

Sensitivity answers the question: **If an individual has a disease or risk factor, what is the probability that the result of the test applied is positive?**

This can be calculated using the following **formula**:

$$\text{Sensitivity} = \text{True positives} / (\text{true positives} + \text{false negatives})$$

or

$$\text{Sensitivity} = a / (a + c)$$

Interpretation of Sensitivity Values

- » A test with 100% sensitivity correctly identifies all patients with the disease.
- » A test with 80% sensitivity correctly detects 80% of patients with the disease (**true positives**) but 20% with the disease go undetected (**false negatives**).

Specificity

Specificity refers to the ability of a test to correctly identify those patients without the disease.

Specificity answers the question: **If an individual does not have the disease or the risk factor, what is the probability that the test applied is negative?**

This can be calculated using the following **formula**:

$$\text{Specificity} = \text{True negatives} / (\text{false positives} + \text{true negatives})$$

or

$$\text{Specificity} = d / (b + d)$$

Interpretation of Specificity Values

- » A test with 100% specificity correctly identifies all patients without the disease.
- » A test with 80% specificity correctly detects 80% of patients without the disease (**true negatives**) but 20% patients without the disease are incorrectly identified as patients with the disease (**false positives**).

Both sensitivity and specificity are **intrinsic characteristics** of the **test itself**. That means that if the test is applied in a population of similar individuals, those values will not vary when it is used in different studies.

Predictive Values

In clinical practice, when a physician requests a diagnostic test, it is unknown if the patient has the disease. That's why physicians should make inferences about the presence or absence of the disease based on the test results. **Predictive values** are a way to quantify this inference.

Positive Predictive Value

The **positive predictive value** (PPV) is the probability that an individual with a positive result truly has the disease.

It is calculated using the following **formula**:

$$\text{PPV} = \text{True positives} / (\text{true positives} + \text{false positives})$$

or

$$\text{PPV} = a / (a + b)$$

The **positive predictive value** answers the question: "How likely is it that this patient has the disease given that the test result is positive?"

Negative Predictive Value

The **negative predictive value** (NPV) is the probability that an individual with a negative result truly does not have the disease.

It is calculated using the following **formula**:

$$\text{NPV} = \text{True negatives} / (\text{true negatives} + \text{false negatives})$$

or

$$\text{NPV} = d / (c + d)$$

The **negative predictive value** answers the question: "How likely is it that this patient does not have the disease given that the test result is negative?"

The prevalence of the disease is the **most determining factor** of predictive values.

Predictive values depend not only on sensitivity and specificity, but also on the **prevalence of the disease**:

- » When the prevalence is **high**, a positive result tends to confirm the presence of the disease, while if it is negative, it will not help to exclude it.
- » When the prevalence is **low**, a negative result will allow the disease to be ruled out with a high margin of confidence, while if it is positive, it will not confirm its existence.

Using Sensitivity and Specificity to Determine the Probability of Disease

Figure 24.1 summarizes data from a hypothetical study of 100 patients presenting pulmonary hypertension.

As a clinician-in-training, I hope you know that tricuspid regurgitation is a complication of pulmonary hypertension and so, you may wonder how accurately a single physical sign (e.g., the presence of a holosystolic murmur at the left lower sternal border) detects this complication.

		Significant tricuspid regurgitation		
		Present	Absent	
Holosystolic murmur	Present	22 a	3 b	25
	Absent	20 c	55 d	75
		42 n ₁	58 n ₂	

Figure 24.1. Summary of data for a hypothetical study of patients with tricuspid regurgitation.

In this hypothetical study, 42 patients have significant tricuspid regurgitation (column 1) and 58 patients do not (column 2).

- » The **sensitivity** of the holosystolic murmur is the proportion of patients with disease (i.e., tricuspid regurgitation, 42 patients) who have the characteristic murmur (i.e., the positive result, 22 patients), which is $22/42 = 0.52$ or 52%.
- » The **specificity** of the holosystolic murmur is the proportion of patients without disease (i.e., no tricuspid regurgitation, 58 patients) who lack the murmur (i.e., the negative result, 55 patients), which is $55/58 = 0.95$ or 95%.

To recall how to calculate sensitivity and specificity, Sackett et al, have suggested helpful **mnemonics**:

- » **Sensitivity** is “pelvic inflammatory disease” (or “PID,” meaning “positivity in disease”).
- » **Specificity** is “National Institutes of Health” (or “NIH,” meaning “negativity in health”).

Figure 24.1 can be used to determine the **accuracy** of the holosystolic murmur, that is, how well its presence or absence discriminates between those with tricuspid regurgitation and those without it. Of the 25 patients who have the murmur (i.e., the positive results), 22 have tricuspid regurgitation; therefore, the probability of tricuspid regurgitation, if the murmur is present (positive finding), is $22/25$ or 88% (i.e., the “post-test probability” if the murmur is present).

Of the 75 patients without the murmur, 20 have tricuspid regurgitation; therefore, the post-test probability of tricuspid regurgitation, if the murmur is absent (i.e., negative finding) is $20/75$ or 27%.

Key Terms

Define the following terms.

2x2 contingency table
False negative
False positive
Gold standard

Negative predictive value
Positive predictive value
Predictive value
Sensitivity

Specificity
True negative
True positive
Validity

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. You have two different tests (A and B) to diagnose a disease. Test A has a sensitivity of 98% and a specificity of 80%, while test B has a sensitivity of 75% and a specificity of 99%.

- Which test is better to confirm the disease?
- Which test is better to rule out the disease?
- Often a test is used to discard the presence of the disease in a large amount of people apparently healthy. This type of test is known as screening test. Which test will work better as a screening test?
- How would you combine both tests to have a higher confidence in the diagnosis of the disease? Calculate the post-test probability of having the disease with the combination of both tests, if the outcome of both test is positive for a prevalence of 0.01.

2. A screening test for lead toxicity in children has a reported sensitivity of 95% and specificity of 80%. The test is administered to 1000 children from a neighborhood in which the prevalence of lead toxicity is 2%.

- Draw a 2x2 contingency table and determine how many children who do not have actual lead toxicity would be expected to test positive (false positive cases).

- Imagine that the same test is administered to 1000 children from a neighborhood in which the prevalence of lead toxicity is 50%. Calculate the probability that a child from this population tests negative given the fact that he/she truly does not have lead toxicity.

3. A screening test for a preclinical stage of a cancer is known to have a sensitivity of 0.90 and a specificity of 0.96. The prevalence of this cancer in its preclinical phase in the population is 1 per 1000 (.001). Assume you use this test in one-hundred thousand (100,000) people. Based on this information, determine:

- Total patients with disease.
- Number of true positives.
- Number of true negatives.
- Number of false negatives.
- Number of false positives.
- Total number of test positives.

4. A screening test for a newly discovered disease is being evaluated. In order to determine the effectiveness of the new test, it was administered to 880 workers, and 120 of the individuals diagnosed with the disease tested positive. A negative test finding occurred in 50 people who had the disease.

A total of 40 persons not diseased tested positive for it. Construct a 2×2 table and calculate the following:

- Prevalence of the disease.
- Sensitivity.
- Specificity.
- Positive predictive value.
- Negative predictive value.
- Likelihood ratio positive.
- Likelihood ratio negative.

5. As an occupational health epidemiologist, you are required to measure the effect of stress on the workers in your manufacturing plant. Two different tests previously developed to measure stress in industrial workers are selected: stress test alpha and stress test delta. Their sensitivities (Sen) and specificities (Spe) are:

Alpha: Sen = 60%; Spe = 95%.
Beta: Sen = 75%; Spe = 90%.

- Which test generates the greatest proportion of false negatives?
- Which test generates the greatest proportion of false positives?
- Which test would you prefer?

6. Multiple choice questions.

1. For a clinical trial, the Sensitivity is = 0.562 and Specificity is = 0.893. This means that:

- a) The test is a valuable test because both indicators are more than 50%.
- c) The test is a worthless test, since it gives errors when detecting both sick and healthy subjects.
- c) The test is a worthless test, because the sensitivity is too low (lower than 75%).
- d) A perfect test.

2. A man presents to his primary care physician complaining of low-grade fevers, diarrhea, and a 15 lb weight loss over the past 3 months. He has used intravenous drugs for 10 years. Physical examination reveals enlarged cervical and femoral lymph nodes. Based on the history and physical examination findings, you estimate that this patient has an approximate 30% pretest probability of having HIV disease and order the HIV antibody test. If the test comes back positive, what is the probability that this patient has HIV disease?

- a) 11%.
- b) 18%.
- c) 41%.
- d) 76%.
- e) 97%.

3. Which one of the following statements about a diagnostic test is true?

- a) The sensitivity of the test is equal to the proportion of individuals without the disease who are correctly identified by the test.
- b) The specificity of the test is equal to the proportion of individuals with the disease who are correctly identified by the test.
- c) The positive predictive value of the test is the proportion of individuals with a positive test result who have the disease.
- d) The positive predictive value of the test is the proportion of individuals with the disease who are correctly identified by the test.
- e) The likelihood ratio for a positive test result is the chance that the patient has the disease if he or she tests positive for it divided by the chance that the patient has the disease if he or she tests negative.

4. A high sensitivity for a diagnostic test means that:

- a) Your doctor is tuned into your needs.
- b) If your test is positive, you have a disease;
- c) The “true-positive” rate for the test is high.
- d) You probably do not have the disease.

5. A high specificity for a diagnostic test means that:

- a) The test is specific for the diagnosis of the disease.
- b) Most people who are test-negative do not have the disease.
- c) Most people who do not have the disease are test-negative.
- d) A positive result indicates that you probably do have the disease.

6. If the positive predictive value (PPV) is high, a positive test result means that:

- a) The disease is almost certainly present.
- b) The disease is more likely to be present if the PPV was calculated from a random population sample.
- c) You cannot be certain that the disease is absent.
- d) Another test is needed to decide the presence of the disease.

7. If the negative predictive value (NPV) is low, a negative test result means that:

- a) Another test is needed to rule out the disease.
- b) You almost certainly do not have the disease.
- c) You cannot be certain that the disease is present.
- d) There is a low probability that the disease is absent.

8. Based on all the information currently available, you estimate that the patient in your office has a one in four chance of having a serious disease. You order a diagnostic test with sensitivity of 95% and specificity of 90%. The result comes back positive. The chance your patient really has the disease is closest to

- a) 30%.
- b) 60%.
- c) 75%.
- d) 90%.

9. The sensitivity of dyspnea on exertion for the diagnosis of coronary heart failure is 100% and the specificity 17%. A negative result implies:

- a) Rules in the disease coronary heart failure.
- b) Rules out the disease coronary heart failure.
- c) Statistical insignificance.
- d) The test is not evaluated in a wide spectrum of patients.

10. A test with 99.9% sensitivity and 99% specificity is used to screen a population for a disease with 1% prevalence. The proportion of test positives in the screen who actually have the disease will be roughly:

- a) 10%.
- b) 30%.
- c) 50%.
- d) 90%.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ in Anaesth, Crit Care and Pain* 2008; 8: 221-223.
- McGee S. Diagnostic Accuracy of Physical Findings. In: McGee S. Evidence-Based Physical Diagnosis. Philadelphia: ELSEVIER; 2012.
- McNamara LA, Martin SW. Principles of Epidemiology and Public Health. In: Long SS, Prober CG, Fisher M. Principles and Practice of Pediatric Infectious Diseases. 5ª Edition. Philadelphia: ELSEVIER; 2018.
- Parfrey PS, Barret BJ. Clinical Epidemiology. Practice and Methods. 2ª Edition. New York: Springer; 2015.
- Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008; 56(1): 45–50.
- Ting K. Sensitivity and Specificity. In: Sammut C., Webb GJ. Encyclopedia of Machine Learning. Springer, Boston, MA; 2011.
- Thelle DS, Laake P. Epidemiology. In: Laake P, Benestad HB, Olsen BR. Research in Medical and Biological Sciences. From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.
- Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front Public Health*. 2017; 5: 307.

ROC Curves

Learning objectives for this chapter

- A. Define what a ROC curve is.
- B. Define terms used in ROC curves analysis.
- C. Determine the relation between sensitivity, specificity, and threshold with ROC curves.
- D. Understand the basic principles of ROC curves analysis.
- E. Identify the application of ROC curves in the clinical practice.

A **Receiver Operating Characteristic (ROC) curve** is a graph that displays the relationship between the **true positive rate** (on the vertical axis) and the **false positive rate** (on the horizontal axis).

Brought into the medical field from engineering usage, a ROC curve helps to **choose the critical value** at which a predictor **best discriminates between choices**.

As we reviewed in the last Chapter, **sensitivity** and **specificity** constitute the basic measures of performance of diagnostic tests. When the results of a test fall into one of two obviously defined categories, such as either the presence or absence of a disease, then the test has only one pair of sensitivity and specificity values. However, in many diagnostic situations, making a decision based on a **binary mode** is both difficult, impractical, and insufficient to describe the full range of diagnostic performance of a test.

To deal with these multiple pairs of sensitivity and specificity values, you can draw a graph using the **sensitivities** as the “y” coordinates and the **1-specificities** or **False Positive Rate (FPR)** as the “x” coordinates. Each discrete point on the graph, called an **operating point**, is generated by using different cutoff levels for a positive test result. The graph obtained is called a receiver operating characteristic (ROC) curve.

The **performance** of a diagnostic variable can be quantified by calculating the area under the ROC curve (**AUROC**). AUROC, is interpreted as the **average value of sensitivity for all possible values of specificity**; is a measure of the overall performance of a diagnostic test.



Bridge to Sensitivity, and Specificity

Sensitivity: the ability of a screening test to detect a true positive, reflecting a test's ability to correctly identify all people who have a condition.

Specificity: the ability of a screening test to detect a true negative, reflecting a test's ability to correctly identify all people who do not have a condition.

AUROC can take on any value between **0 and 1**, where a bigger value suggests the **better overall performance** of a diagnostic test. The ideal test would have an AUROC of 1, whereas a random guess would have an AUROC of 0.5.

ROC Curves at a Glance

An **ideal test** would have sensitivity and specificity both equal to 1. If a cut-off value existed to produce such a test, then the sensitivity would be 1 for any non-zero values of $1 - \text{specificity}$. The ROC curve would start at the origin (0,0), go vertically up the y-axis to (0,1) and then horizontally across to (1,1). A good test would be somewhere close to this ideal (**Figure 25.1**).

If a variable has no **diagnostic capability**, then a test based on that variable would be equally likely to produce a false positive or a true positive:

- » Sensitivity = $1 - \text{specificity}$, or
- » Sensitivity + specificity = 1.

This equality is represented by a diagonal line from (0,0) to (1,1) on the graph of the ROC curve, as shown in **Figure 25.1** (dashed line).

When a test is not that perfect (like most of the times), both distributions **overlap**. Depending upon the threshold, we can minimize or maximize the appearance of **Type I** and **Type II errors** (**Chapter 10**). For example, **Figure 25.2** shows a scenario where the AUROC is 0.7, meaning that there is 70% chance that model will be able to distinguish between true positives and true negatives.

What happens when a test **isn't perfect at all**? This is the worst situation. **Figure 25.3** shows a scenario where AUROC is approximately 0.5, model has **no discrimination capacity** to distinguish between true positives and true negatives.

When the AUROC is approximately 0, the model is actually **reciprocating the classes**. It means, the model is predicting true negatives as true positives and vice versa (**Figure 25.4**).

ROC Curves as a Tool to Help Choosing Between Diagnostic Tests

The ability of **two continuous variables** to diagnose an outcome can be compared using ROC curves and their AUROCs. The decision to use a diagnostic test depends not only on the ROC analysis but also on the **ultimate benefit to the patient**, and the prevalence of the outcome (pre-test probability). Generally, there is a **trade-off** between sensitivity and specificity, and the practitioner must make a decision based on their relative importance.

Statistical
power

Type I error: rejecting the null hypothesis when it is in fact true.

Type II error: not rejecting the null hypothesis when it is in fact not true.



Bridge to

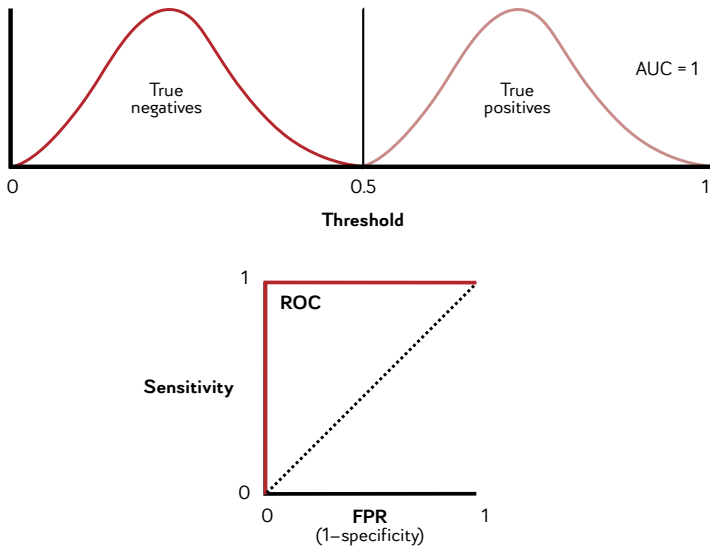


Figure 25.1. Graphic representation of a model of a perfect diagnostic test and its ROC curve.

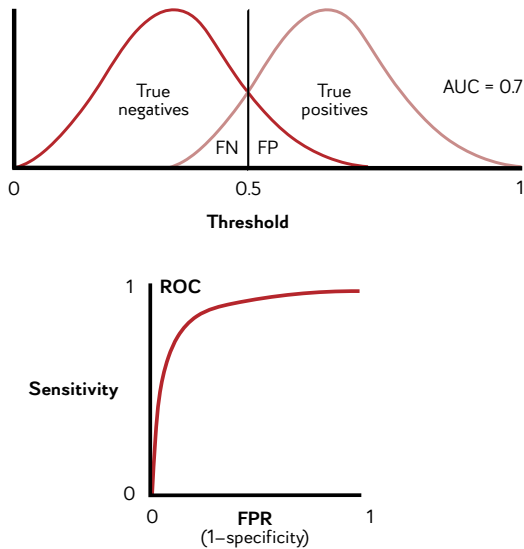


Figure 25.2. Graphic representation of a model of a less perfect diagnostic test and its ROC curve.

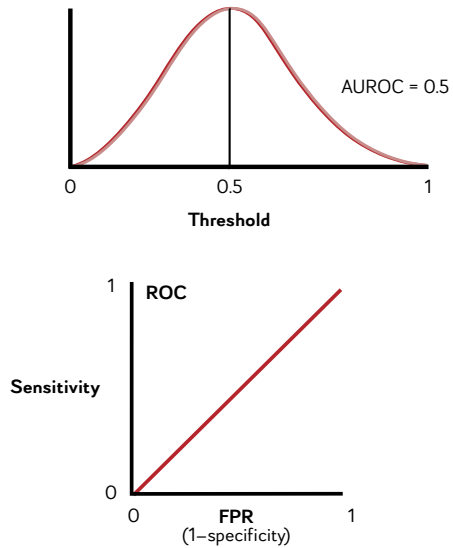


Figure 25.3. Graphic representation of a model of an imperfect diagnostic test and its ROC curve.

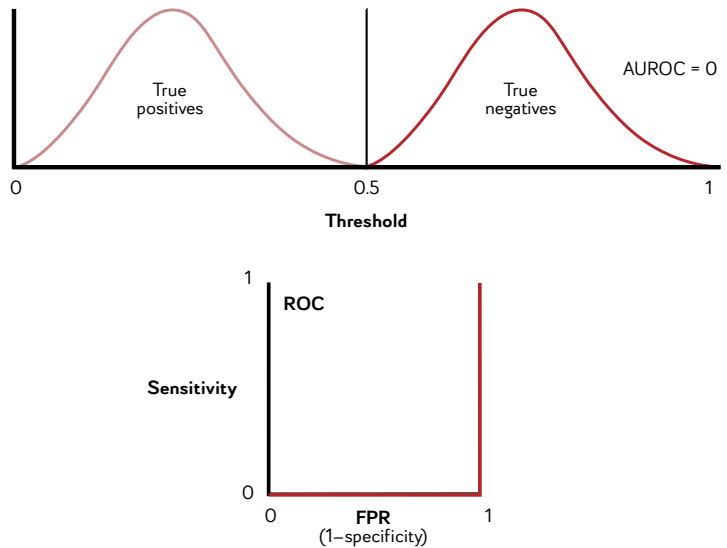


Figure 25.4. Graphic representation of a model of a reciprocal diagnostic test and its ROC curve.

Key Terms

Define the following terms.

Area under the ROC curve

Discrimination capacity

Receiver operating characteristic (ROC) curve

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. When dealing with ROC curves:

- What do you graph in the “y” coordinates?
- What do you graph in the “x” coordinates?

2. Define AUROC in your own words.

3. Find a ROC curve in the literature and interpret the true positive rate and the false positive rate.

4. Multiple choice question.

1. Urinalysis provides a relatively costly but accurate test for proteinuria (i.e. when the albumin–creatinine ratio (ACR) is greater than or equal to 30 mg/g). In 2011, White et al. investigated the ability of a cheaper urine dipstick test to identify those with proteinuria. Based on data from 10944 individuals with complete urinalysis data, White et al. investigated the ability of an ACR value ≥ 30 mg/g measured on a urine dipstick to identify individuals with ACR ≥ 30 mg/g based on urinalysis. The area under the receiver operating characteristic curve (AUROC) for dipstick detection of ACR ≥ 30 mg/g was 0.85 in men and 0.78 in women. Which one of the following statements is true?

a) Among men, an ACR of ≥ 30 mg/g from a urine dipstick will correctly identify 85% of those who really do have proteinuria.

b) The AUROC of 0.78 in women indicates that 78% of those with a raised ACR on a dipstick

will truly have proteinuria based on urinalysis.

c) The AUROC is calculated as sensitivity divided by (1 minus specificity).

d) Given two randomly selected women from the sample, one of whom does and one of whom does not have ACR ≥ 30 mg/g based on dipstick analysis, the woman with the ACR ≥ 30 mg/g based on dipstick analysis has a 78% probability of also having an ACR ≥ 30 mg/g based on urinalysis.

e) Given two randomly selected women from the sample, one of whom does and one of whom does not have proteinuria based on urinalysis, the dipstick analysis will correctly identify the woman with proteinuria on 78% of occasions.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. *Crit Care*. 2004; 8(6): 508–512.
- Concejero P. Comparación de modelos de curvas ROC para la evaluación de procedimientos estadísticos de predicción en investigación de mercados. Tesis Doctoral. Universidad Complutense de Madrid; 2004.
- Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*. 2013; 4(2): 627–635.
- Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*. 2005 Oct;38(5):404-15.
- Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol* 2018;29(7):07TR01.
- Park SH, Goo JM, Jo CH. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J Radiol*. 2004 Jan-Mar; 5(1): 11–18.
- Riffenburgh RH. Risks, Odds, and ROC Curves. In: Riffenburgh RH. *Statistics in Medicine*. 3ª Edition. Burlington: Elsevier Elsevier Academic Press; 2012.

Section VI

Clinical Interventions

Chapters of the Section

Chapter 26	Introduction to Experimental Study Designs
Chapter 27	Ethics in Clinical Trials
Chapter 28	Randomized Clinical Trials (RCT)
Chapter 29	Equivalence, and Non-Inferiority Trials
Chapter 30	Efficacy, Effectiveness, Efficiency
Chapter 31	Bias

Introduction to Experimental Study Designs

Learning objectives for this chapter

- A. Identify the main characteristics of an experimental study.
- B. Identify the main objectives of experimental studies.
- C. Identify the experimental studies as the most powerful study design.

Experimental studies are those in which the research team **controls the study factor**, that is, the team decides which subjects will receive the intervention to be evaluated, as well as how they will do it (dose, schedule, duration, etc.), according to a pre-established research protocol. This means that, by definition, experimental studies are **prospective**.

The **main objective** of experimental studies is to evaluate the **effects** of an intervention, trying to establish a **cause-effect relationship** with the observed results (they are, therefore, **analytical** studies). This intervention is usually a pharmacological treatment, although it can be any other type of therapy, a health council, a preventive activity, a diagnostic strategy or an organizational model.

Given that the intervention is administered to the subjects for the purpose of study, the **ethical requirements** of research in human beings are especially important, and only potentially beneficial interventions for subjects should be evaluated. These interventions must have sufficient prior information to justify the performance of the experiment, and the studies must be designed in accordance with the accepted scientific standards, both ethical and methodological.

The great **advantage** of experimental studies over observational studies is that, by controlling the study factor as well as the conditions under which the research is carried out, it reduces the possibility that other factors may influence the results. This provides greater confidence in the results obtained (**higher quality of the evidence**).

Experimental designs are the **most powerful study designs** because the investigator controls the exposure or treatment.

This enables to make the comparisons between groups as similar as possible, so that any differences observed should be attributable to the exposure or treatment.

The most important experimental design is the **randomized clinical trial** (RCT), whose general characteristics are described in a following chapter.

Key Terms

Define the following terms.

Analytical study

Ethical requirements

Experimental study

Prospective study

Randomized clinical trial (RCT)

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Make a comparison list between the characteristics of an experimental study vs. the characteristics of an observational study.
2. Draw a table comparing the pros and cons between an experimental study and an observational study.
3. Randomized clinical trials are the most important experimental design. Search for and list other experimental designs available.
4. Identify the most important characteristics of the architecture of an experimental study. Go back to **Chapter 13. Study Designs** and review their meanings and importance.
5. Get together with some classmates and discuss what would be the steps that, as a researcher, you must follow in order to create an experimental study design.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. N Engl J Med 2016; 375:65-74.
- Brody T. Clinical Trial Design. In: Brody T. Clinical trials. Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. 2ª Edition. Cambridge: ELSEVIER; 2016.
- Evans SR. Fundamentals of clinical trial design. Exp Stroke Transl Med. 2010 Jan 1; 3(1):19-27.

Ethics in Clinical Trials

Learning objectives for this chapter

- A. Define the importance of Ethics in clinical trials.
- B. Identify the main historical events that lead to the consolidation of Ethics in clinical trials.
- C. Identify the main guidelines that determine the role of Ethics in clinical trials in Mexico.
- D. Know the main components of the national legal framework of Mexico.

Clinical research has been conducted over hundreds and even thousands of years. However, the concept and standardization of assuring Ethics in clinical research is much more recent.

This standardization began in Europe after the Holocaust when atrocities in the name of “medical research” were conducted on concentration camp prisoners. These war crimes were the subject of trials that took place in the German city of **Nuremberg** after the war. In 1947, the **Nuremberg Code** was the first written declaration of ethical principles in clinical research, which serves as the basis of today’s standards. Nonetheless, it was only in the late 1970s that the United States took similar actions in standardizing and documenting ethical principles for clinical research, and systematically applying these principles. Now, these various principles are internationally recognized, accepted, and reasonably well harmonized.

The Nuremberg Code

At the end of World War II, in Nuremberg, Germany, 15 Nazis were investigated and convicted for war crimes for conducting unethical “medical research” on prisoners from concentration camps.

The legal decisions written for the court case against the war criminals formed the basis for the “**Nuremberg Code**” (1947).

The **basic principles** outlined in the Nuremberg Code included the following:

- » Participation in clinical research must be **voluntary**.
- » Participation in clinical research must be based on an **informed consent**.
- » Risks should be **minimized**, and the risks should be justified by **anticipated benefits**.
- » Only **qualified researchers** should conduct clinical (medical) research.
- » The clinical research should be justified based on **prior research in animals**.
- » Physical and mental **suffering** should be **avoided**.
- » Research is prohibited if **death or disability** is the expected outcome.

Declaration of Helsinki

In **1964**, the World Medical Association met in Helsinki, Finland. The outcome of this meeting was the **Declaration of Helsinki**. This declaration has been revised several times since, as ethical principles have been refined (elaborated), and more ethical conundrums have been considered.

The additional principles that derived from the 1964 Declaration (besides those previously defined in the Nuremberg Code) included the following:

- » Precedence is given for **well-being of trial subject** (over research being performed or investigator's interests)—i.e., the highest priority in conducting research is **assuring the rights and well-being of the trial subject**.
- » **Respect** for persons should be implicit.
- » **Protection** of subjects' **health** and **rights** should be assured.
- » Special protections for **vulnerable** populations should be afforded to them.

The Belmont Report

In the United States, physicians and medical researchers felt that their practices were already ethical. Nonetheless, In 1966, Beecher outlined 22 examples of unethical research taking place in the U.S. and in Europe.

In 1974, a long, hard look at research practices in the United States was taken, when the U.S. Congress authorized the creation of the **National Commission for Protection of Human Subjects in Biomedical and Behavioral Research** (referred to as **National Commission**). As an outcome, ethical principles emerged to regulate biomedical and behavioral research in the United States.

This National Commission published the **Belmont Report** in 1979, which contained the guides and ethical considerations of clinical research in the U.S. Furthermore, the Belmont Report distinguished clinical research from the “practice of medicine”.

The Belmont Report summarized ethical principles into **three categories of considerations**, as follows:

1. **Respect for Persons.**
2. **Beneficence.**
3. **Justice.**

Respect for Persons

The **principles** that guide this consideration include:

- » Individuals should be treated as **autonomous agents**.
- » Individuals with diminished autonomy should be afforded **additional protections** (for vulnerable populations).

Translated into **practice**, the application of this principles, means:

- » Informed consent.
- » Respect for subjects’ privacy.
- » Additional protections for vulnerable populations.

Beneficence

The **principles** that guide this consideration include:

- » Do not harm (non-maleficence).
- » Maximize potential benefits and minimize potential risks

Translated into **practice**, the application of this principles, means:

- » Need for assessment of risks and benefits.
- » Increased benefits and decreased harms in a manner that is consistent with sound research design (balance between needs of rigorous science with the ethical principles protecting the trial participants).

The Justice principle assures equitable enrollment of clinical research subjects into the research study.

- » Requirement for researchers to be qualified to perform the research.
- » Prohibition of research that does not have potential benefits that outweigh potential harm.

Justice

The **principle** that guides “justice” is **fairness** in who derives benefits from the research and who bears the burdens of it.

Translated into **practice**, the application of this principle, means:

- » Selection of subjects for enrollment by fair inclusion/exclusion criteria.
- » Avoidance of exploitation of vulnerable populations or populations of convenience.

Ethics in Mexico

The Declaration of Helsinki is the main international document that guides medical research around the world, so it is essential to take it into account when designing and developing research that involves human beings.

Mexico is not the exception. Those who carry out research for human health in our country must adapt to the scientific and ethical principles set forth in that Declaration, as well in the internationally accepted instruments mentioned, and also take into consideration what the **Comisión Nacional de Bioética (CONBIOÉTICA)** declares.

The CONBIOÉTICA promotes the establishment and operation of **Hospital Bioethics Committees and Research Ethics Committees** in public and private health institutions in Mexico. It also establishes, through the National Guide for the Integration and Operation of the Research Ethics Committees, the criteria for the development of activities and training of members of the collegiate bodies.

In order to promote strengthening and regulating medical research, in 2012, the Norma Oficial Mexicana (**NOM-12-SSA3-2012**) was created, which determines the criteria for the execution of research projects for health in human beings.

These considerations are **administrative, ethical and methodological** in nature, and are coordinated with the provisions of the Ley General de Salud, and the Regulations on health research.

National Legal Framework

In Mexico, the legal framework for health has been transformed in recent decades.

The most relevant normative **instruments** in the study of bioethical and research ethics issues are listed below.

- » Political Constitution of the United Mexican States.
- » Organic Law of the Federal Public Administration.
- » General Health Law.
- » General Law of Transparency and Access to Public Information.
- » Federal Law on Transparency and Access to Government Public Information.
- » Federal Law on the Protection of Personal Data Held by Individuals.
- » Federal Law of Administrative Procedure.
- » Regulation of the General Health Law in Health Research Matters.
- » Regulation of the General Health Law on the Provision of Care Medical Services
- » Regulations of the Federal Commission for Protection against Health Risks.
- » Internal Regulations of the Ministry of Health.
- » Decree creating the decentralized body called the National Commission for Bioethics.
- » Agreement by which the General Provisions for Integration and Functioning are issued and established in conformity with the criteria established by the National Bioethics Commission.
- » Agreement by which the diverse one by which the General Provisions for the Integration and Operation of Research Ethics Committees are issued and is amended and the hospital units that must have them are established, in accordance with the criteria established by the National Bioethics Commission, published on October 31, 2012.
- » Agreement that establishes the guidelines that must be observed in public establishments that provide health care services to regulate their relationship with manufacturers and distributors of medicines and other health supplies, derived from the promotion of products or the conduct of academic activities, research or scientific.

Key Terms

Define the following terms.

Belmont Report
Beneficence
CONBIOÉTICA

Declaration of Helsinki
Justice
NOM-12-SSA3-2012

Nuremberg Code
Respect for Persons
Legal framework

Active–Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- 1. State at least 3 internationally recognized documents that are necessary to justify conducting a medical experiment in humans.**
- 2. State the basic principles outlined in the Nuremberg Code.**
- 3. State the principles derived from the 1964 Declaration of Helsinki.**
- 4. Draw a diagram of the considerations in The Belmont Report with their principles and clinical applications.**
- 5. Answer:** what is the main international document that guides medical research around the world?
- 6. Answer:** what is the official document in Mexico that determines the criteria for the execution of research projects for health in human beings?
- 7. Answer:** what is the main normative instrument in the study of bioethical and research ethics issues in Mexico?

Bibliography and Suggested Reading

- Argimon–Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Comisión Nacional de Bioética. Guía nacional para la integración y el funcionamiento de los Comités de Ética en Investigación. 5ª Edición. México: Secretaría de Salud/Comisión Nacional de Bioética; 2016.
- Lara-Gutiérrez YA, Pompa-Mansilla M. Ética en la investigación en educación médica: consideraciones y retos actuales. Revista Investigación en Educación Médica. 2018;26(7):99-108.
- Miranda AG, Torres FC. Ética en la investigación médica. Rev. Soc. Andaluza Traumatol Ortop. 2008; 26:119-22.
- Norma Oficial Mexicana NOM-012-SSA3-2012. 2012.
- Rajendran N, Thomas D, Madhavan SS, Herman RA. Ethics in Clinical Research. In: Thomas D. Clinical Pharmacy Education, Practice and Research. Clinical Pharmacy, Drug Information, Pharmacovigilance, Pharmacoeconomics and Clinical Research. Amsterdam: ELSEVIER; 2019.
- Sheets R. Clinical Trial Ethics, Human Subjects Protections, and the Informed Consent Process. In: Sheets R. Fundamentals of Biologicals Regulation. Vaccines and Biotechnology Medicines. United Kingdom: ELSEVIER; 2018.
- World Health Organization. Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants. WHO Document Production Services: Geneva, Switzerland; 2011.
- World Medical Association: WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Beings. Available at: <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>.

Randomized Clinical Trials (RCT)

Learning objectives for this chapter

- A. Identify the RCT as the most important experimental study design.
- B. Identify the interventions that can be compared with a RCT.
- C. Learn about the different options for selecting a reference intervention.
- D. Understand the classification of RCTs.
- E. Understand the methodological design of RCTs.
- F. Define study validity.
- G. Differentiate between internal and external validity and understand their importance in RCTs.
- H. Identify and understand the critical issues in the design of a RCT.
- I. Identify the characteristics that must be met to select the population to be included in a RCT.
- J. Determine the importance of randomization in RCTs.

Randomized Clinical Trials (RCTs), when correctly designed and rigorously conducted, provide the **most definitive answers** regarding intervention effects, serve as the **standard for clinical research**, and have contributed immensely to advances in patient care. Nevertheless, other clinical trial designs and observational investigations can be appropriately employed depending on resources and the specific question of interest.

As mentioned in a previous Chapter, the **essential purpose of the RCT** is to determine whether a particular intervention or treatment can reasonably be inferred to cause a change in health, disease progression, or risk factor(s) associated with a disease. As such, it is considered the **best available design to evaluate the efficacy** of a health intervention.

For this reason, therapeutic recommendations and **clinical practice guidelines** are being constructed, with increasing frequency, based on the evidence provided by this type of study.

When trying to evaluate **efficacy**, a **RCT** is the **ONLY** study design that can fulfill that quest.

This has led to the rapidly **increase** in the number of RCTs performed and to a greater methodological rigor in its design, execution, and analysis. In addition, it has led to the development of instruments for assessing its methodological quality, guides for quick reading and recommendations on its publication.

Table 28.1 summarizes the advantages and disadvantages of the RCTs, and **Figure 28.1** shows the basic structure of a RCT.

Interventions Compared in a RCT

One of the key aspects of RCT design is the selection of the **intervention** to be used as a **reference** in the comparison.

The nature of the interventions could be **pharmacological, surgical, behavioral, device, strategy-based** or could consist of **multiple components**.

No matter what the intervention is, the so-called uncertainty principle (**Equipoise**) must be respected. This means that an RCT should only be performed if, in light of the available evidence, both interventions being compared can be considered **reasonably therapeutic alternatives for patients**, since there are doubts about whether one of them is superior to the other.

Comparison with interventions known to be inferior, in addition to being **ethically unacceptable**, leads to favorable results for the study intervention, whose publication introduces a **bias** in the available evidence on the efficacy of treatments, with the repercussions that this fact can have on decisions and therapeutic recommendations.

Table 28.1. Advantages and Disadvantages of RCTs

Advantages	Disadvantages
<ul style="list-style-type: none"> • They provide the best evidence of a cause-effect relationship between the evaluated intervention and the observed response. • They provide greater control of the study factor. • Random assignment produces a balanced distribution of prognostic factors that can influence the result (potential confounding factors), so that comparable groups are formed, and the effect of the intervention is isolated from the other factors. 	<ul style="list-style-type: none"> • Ethical constraints prevent many questions from being addressed in a RCT. • They are usually carried out with highly selected participants, making it difficult to generalize and extrapolate the results. • Often, interventions are administered with rigid guidelines, different from those carried out in routine practice, making it difficult to generalize and extrapolate the results. • They only allow evaluating the effect of a single intervention. • They usually have a high cost, although it depends on the duration of the study and the complexity of the protocol.

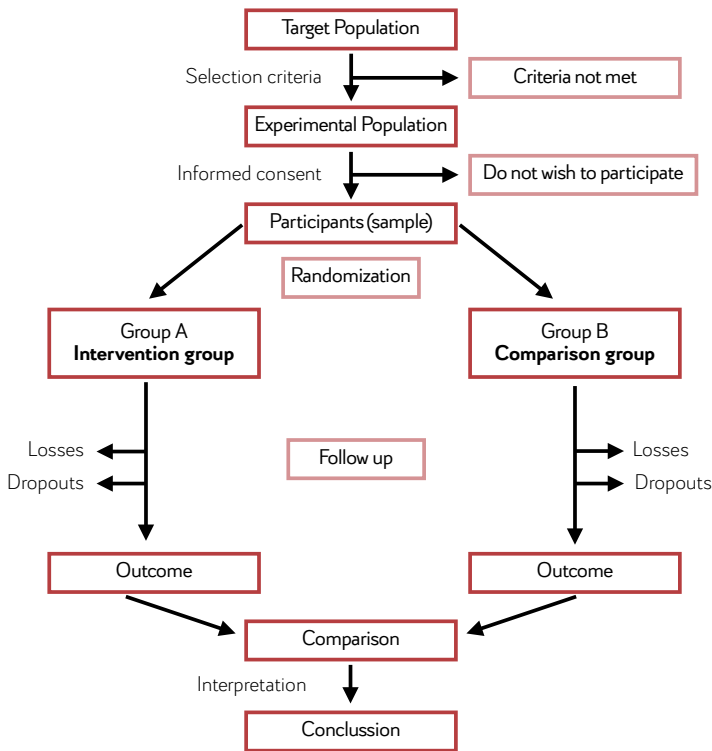


Figure 28.1. Basic structure of a randomized clinical trial.

In general terms, there are **three options** for selecting the reference intervention: **placebo**, **active treatment** or **no intervention**.

Placebo

A **placebo** is a preparation that lacks pharmacological activity, but whose appearance and organoleptic characteristics are identical to those of the study preparation.

The purpose of its use in research is to **control the placebo effect**, which refers to the psychological or physiological effect of any medication, regardless of its pharmacological activity, and which depends on factors such as the patient's own personality, the convictions and the enthusiasm of the research team, the conditions of administration and the characteristics of the intervention, etc.

If an anti-hypertensive is compared with a placebo, it is being evaluated whether the treatment manages to lower the blood pressure levels beyond what a pharmacologically non-active substance would.

The comparison with a placebo is intended to **quantify the therapeutic effect of the drug being evaluated**, since the main advantage of placebo as a comparison alternative is control the effects derived from any characteristic of the treatment other than the effect being studied, including side effects.

Another Active Treatment

The objective of using another active treatment or intervention as a comparison is to **estimate the benefit-risk ratio** of the new treatment in a specific clinical situation.

In these cases, the best comparison is the “**best available treatment**” for that situation.

» This is not always an easy choice, since in most cases there is a wide therapeutic arsenal that makes it difficult to choose which is the best alternative.

When an active treatment is used as a reference, special attention should be paid to the **dose**, the **regimen** and the **duration** of its administration, to maintain the **uncertainty principle** and not favor the new treatment when comparing it with another inferior treatment.

When comparing two active treatments, it is often desirable to **mask** them to prevent possible **bias**.

» To achieve this, it is necessary to administer to each group a placebo of the intervention received by the other group (**double dummy**). In this case, the placebo is not used as a reference alternative, but only as a technique to achieve masking.

No Intervention

Sometimes, because of the research question itself, the most appropriate reference intervention is not a placebo nor another specific intervention, but rather the **usual care** that patients receive in the office.

Although it is possible to compare the group that receives the study intervention with another that does not receive any specific intervention, it can be considered that the control group is receiving the care that is usually provided for their health (otherwise ethical issues might arise), so it is actually being compared to usual care.

Determining the Outcome in a RCT

The choice of the outcome variable to be used to evaluate the **efficacy** of the treatment, quantify its effects, and compare them with those of the reference group is **key** to establishing the clinical relevance of the results to be obtained.

Quality of life or recovery of a given functional capacity, for example, are relevant variables for patients. Many trials focus exclusively on the potential benefits of a treatment, but do not measure its effects on other important variables or even some side and adverse effects.

The **most appropriate variable** should be selected, that is:

- » The one that measures the true results of importance and interest to patients, and not simply because it is easy to measure or because it is expected that it can show changes or differences in a short period of time.

Clinical Trials Classification

Clinical trials are classified into **phases** based on the objectives of the trial.

Phase I Studies

- » Are the first studies of an intervention conducted in humans.
- » They include dose-ranging and safety studies, and traditionally (but not always) are non-randomized.
- » Have small sample sizes (e.g., <20).
- » The fundamental goal of these studies is to investigate **pharmacokinetics**, **pharmacodynamics**, and **toxicity** of the investigational intervention or treatment.
- » When feasible, a dose-limiting toxicity (DLT) threshold or physical event must be defined to create a **stopping rule**.
 - For example, the appropriate dose of a particular psychotherapy for the treatment of major depressive disorder is important to determine. Such a dosing study needs to establish safety, stopping rules, and specific definition of toxicity for that particular intervention.

Phase II Studies

- » The purpose of phase II studies is to prove that a new therapy has **sufficient activity** (e.g., reduction in a clinically relevant biomarker) to be tested in a larger RCT.
 - Early Phase II studies (also known as **Proof of Concept Studies**) provide the necessary signal that some activity is occurring to justify larger and more expensive trials.
- » Randomized or not, early Phase II designs typically require a relatively **small number of patients**.
 - This prevents large numbers of patients from being exposed to useless or even potentially harmful treatments when the evidence shows that the benefits of the new therapy are small or nonexistent.
- » The **disadvantage** of the non-randomized and open-label strategies is that there is less experimental control imposed than is optimal.

- Thus, **internal validity is sacrificed** (more on this at “Study Validity” below) and can result in prominent placebo effects, investigator and other sorts of bias due to lack of masking, regression to the mean, and other threats to internal validity.
- » Phase II trials are typically conducted to investigate a **dose-response relationship**, identify an **optimal dose**, and to investigate **safety issues**.

Phase III Studies

- » Phase III studies, which can be **efficacy** or **effectiveness** studies (more on this in a following chapter), are large **prospective trials** designed to compare an experimental intervention to a control (or standard) intervention.
- » Can be designed to demonstrate **superiority, non-inferiority, or equivalence** (more on this in a following chapter too).
- » They are typically longer in duration than Phase II trials and employ less control on participant characteristics, delivery of the intervention, and characteristics of the study environment.
- » Some are more “real world” than are earlier phase studies.
- » While its internal validity is lower, the **external validity** is much higher (more on this at “Study Validity” below).
- » May be used for many types of investigations, including evaluating an intervention for the purposes of **treatment, prevention, or diagnosis**.
- » Are generally large trials (i.e., many study participants) designed to “**confirm**” the efficacy of an intervention.

Phase IV Studies

- » Phase IV studies are very large, typically conducted by pharmaceutical companies for **post-marketing surveillance studies** of population safety, effectiveness, and generalizability in order to gain broader experience with the intervention.
- » Are designed to study **longer-term effects** of treatments on populations.

Clinical Trial Designs

The most optimal design, analytic strategy, and end points for any research study are dependent, more than anything else, on:

- » What **question** is being addressed.
- » Where that specific question fits on the **spectrum of the research continuum**.

In general, **simple designs** with a targeted and well-characterized question, clearly defined end points, and patient characteristics that will allow a clear and definitive answer, are optimal.

Table 28.2 summarizes the different types of clinical trial designs.

Selection of the Population

The **aim of the study** determines the population in which the study will be made. This population is defined by a **selection criteria** specified a priori (**experimental population**), from which the subjects who finally participate in the trial are selected.

Selection Criteria

Selection criteria identify a population in which, based in the current available knowledge, the interventions being compared could be equally indicated, and therefore potentially beneficial. This implies that subjects should be excluded when one of the alternatives is preferable to the other, and when either intervention is contraindicated or could present serious interactions.

- » The use of **strict inclusion and exclusion criteria** leads to obtaining a homogeneous sample, which increases the **internal validity** of the study. Nevertheless, by moving the study population away from the target population, **limits** its ability to **generalize** the results.
- » The use of **very broad inclusion and exclusion criteria** leads to obtaining a more **representative** sample of the target population, and the possibilities of **generalizing** the results will be greater.
 - Nevertheless, the more heterogeneous the sample, the more difficult to detect a response to the treatment, and a **greater number of individuals** will be required.

Informed Consent

Once it has been verified that a subject meets all the inclusion criteria, and none of the exclusion criteria, the participants must give their **informed consent** to participate in the study before being included in it. In Mexico, this is stated in the **Norma Oficial Mexicana (NOM) 004-SSA3-2012** (From the Medical Records).

This NOM defines the informed consent as the **tangible expression of respect for people's autonomy in the field of medical care and health research**. Informed consent is not a document, it is a continuous and gradual process that occurs between health personnel and the patient and that is consolidated in a document.

Table 28.2. Types of Clinical Trial Designs

Trial design	Characteristics
Crossover Designs	<ul style="list-style-type: none"> • Each participant receives all treatments that are being investigated but at different times • The order in which a participant receives the treatments is randomized • After completing Treatment #1, the patient “crosses over” and receives Treatment #2 • Between treatments is a period called a “washout” when no treatment is delivered • Advantages: each patient serves as his or her own control, reducing between-subject variability, and allowing the detection of smaller effect sizes with reduced sample sizes • Disadvantages: when the treatment being investigated has a sustained effect on the outcome of interest • Useful for episodic conditions (if the episodic nature is somewhat predictable) • Problematic for unstable or progressive conditions (they add variability because changes in disease will be introduced over the course of the study)
Enriched Enrollment Designs	<ul style="list-style-type: none"> • A variant of the crossover design • Useful in studying treatments to which only a minority of patients respond • If the results are not statistically significant in a conventional clinical trial, but an intervention appears effective for subpopulations of patients, a potentially useful strategy is to enter responders into a second prospective comparison trial. If the results of the second trial are statistically significant, this suggests that the patients' initial response was not attributable to chance • May be of interest in treatment intervention studies because they demonstrate limited evidence for a treatment response and may suggest further investigation
Factorial Designs	<ul style="list-style-type: none"> • Each level of a factor (treatment or condition under study) occurs in combination with every level of every other factor • Experimental units are randomly assigned to treatment combinations rather than individual treatments • Classically, each intervention should have independent effects; in other words, there must be no interaction between any of the interventions • Major challenges: (1) meet the independence assumption and (2) choose a sufficiently large sample size to be able to detect meaningful interactions with high power or a good statistical chance of observing an interaction (if it truly is present)

Table 28.2. Types of Clinical Trial Designs (continued)

Trial design	Characteristics
Parallel Groups Designs	<ul style="list-style-type: none"> • Participants are randomized to one of several possible treatments • Interest focuses on comparing the effects of the treatments on a common response or outcome • The effect on the response could be adjusted for baseline measurements of patient characteristics • The double-blind randomized parallel groups design is the “gold standard” to which all other designs should be compared
Sequential Trial Designs and Interim Analyses	<ul style="list-style-type: none"> • The parallel groups are studied not for a fixed period of time but, rather, until either a clear benefit from one treatment group appears or it becomes highly unlikely that any difference will emerge • These trials tend to be shorter than fixed-length trials when one treatment is much more effective than the other treatments
Group-Randomized Trial Designs	<ul style="list-style-type: none"> • Also known as “cluster randomized trials” • The unit of randomization is one of the several types of groups, rather than an individual participant • Such groups might include schools, clinics, worksites, communities, or other units • Group randomization to treatment can be an efficient strategy when an intervention is difficult to implement on an individual level without the risk of contamination, such as interventions that affect environments
Adaptive Treatment Designs	<ul style="list-style-type: none"> • Also called “adaptive intervention”, “stepped-care” or “dynamic treatment designs” • Allow changes in the dose or components of an intervention after the onset of the study, as a function of individual or environmental factors or characteristics • Decision rules are established before the onset of the study regarding the characteristics of interest (e.g., gender, outcome of interest) and how they will determine assignment to specific intervention components or dose • Individuals can be randomly assigned to condition several times • These designs, when carefully constructed, can be efficient and cost-effective and are increasingly used because they allow development of individually tailored treatment strategies

Through informed consent, health personnel inform the competent patient, in **sufficient quality and quantity**, about the following:

- » The nature of the disease and the diagnostic or therapeutic procedure intended to be used.
- » The risks and benefits that it entails.
- » Other possible alternatives.

The written informed consent is the evidence that reflects that the medical personnel have informed the patient and that the patient has understood the information. Therefore, it is the manifestation of the responsible and bioethical attitude of medical or health research personnel, which increases the quality of services and guarantees respect for the dignity and autonomy of people.

Case-control studies

In a case-control study, the comparison group allows to evaluate how factors differentiate those individuals who do and do not have the disease under study.



Choosing the Comparison Group

Comparison groups have one important **purpose**: allows to **evaluate the outcome** that could result **in absence** of the experimental or intervention condition under study in a clinical trial.

In all studies, is important to carefully describe both the experimental (or **case**) and comparison (or **control**) groups.

Comparison groups can be participants randomized to a placebo control, usual care, standard of care, attention control, or alternative treatment.

Control Groups

One of the most difficult issues in clinical trial design is how to choose and design the most appropriate control group for a specific treatment and outcome.

The **purpose** of the control group is to **control potential threats to internal validity** so the dependent variable(s) of interest, rather than any other nontreatment-related factors, can be said to be more likely associated with the active ingredient of the experimental treatment.

The primary driving force for the choice of a control group should be **the specific question being addressed**. Different control group conditions will allow different conclusions to be made. In other words, to choose the most appropriate control group mandates that one has a good grasp of what needs to be controlled in the experimental setting, including how the treatment of interest is defined and what the outcomes of interest are.

Some of the **factors meant to be controlled** by control groups include the following:

- » Expectations (by both patient and provider).

- » Time and attention.
- » Practitioner effects.
- » Social support (from practitioner and other sources).
- » Compensation.
- » Demand or burden.
- » Risk.
- » Disease progression.
- » Nonspecific effects, including contextual effects.

Control groups can take different forms, from wait-list control, placebo control, sham control, time and attention control, and active comparator control groups. There is an enormous importance in choosing the most appropriate control group for any specific study. However, the discussion of the advantages and disadvantages of each of these forms of control groups is outside the scope of this book.

Study Validity

The **validity** of a research study refers to how well the results among the study participants represent true findings among similar individuals outside the study.

“Validity” applies to **all types of clinical studies**, including those about prevalence, associations, interventions, and diagnosis.

The validity of a research study includes **two domains** (Figure 28.2):

- » **Internal validity:** the extent to which a causal conclusion represents the truth in the population under study and, it is not because of methodological errors.
 - It is determined by the degree to which a study minimizes **systematic error** (or “bias” or confounding; **Chapter 31**).
 - Can be increased if careful study planning and adequate quality control and implementation strategies are ensured—including adequate recruitment strategies, data collection, data analysis, and sample size.
- » **External validity:** the validity of generalized (causal) inferences from our policy evaluation. Is the degree to which the conclusions in the study would hold for other persons in other places and at other times.
 - Can be increased by using broad inclusion criteria that result in a study population that more closely resembles real-life patients, and by choosing interventions feasible to apply.

Lack of internal validity implies that the results of the study deviate from the truth, and, therefore, we cannot draw any conclusions; hence, if the results of a trial are not internally valid, external validity is irrelevant.

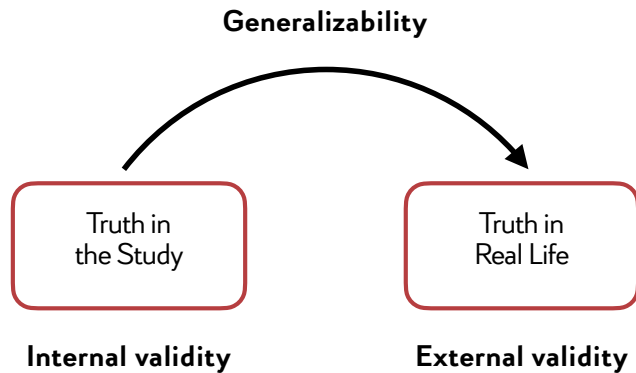


Figure 28.2. Internal and external validity.

Critical Issues in Clinical Study Designs

Blinding or Masking

A longstanding methodological ideal is to keep patients in clinical trials **unaware** of their allocated treatment. Non-blinded patients, aware of their treatment, may differ from blinded patients in how they report symptoms or in the quality of the doctor-patient relationship, inducing dissimilar rates of co-intervention, attrition and placebo effect.

Blinding or masking, when possible, may be **almost as important as randomization** itself.

The different parties involved in a clinical trial are all possible sources of **bias**, and such they can be blinded to ensure **objectivity**. This parties include:

- » The patient being treated.
- » The clinical staff administering the treatment.
- » The physician assessing the treatment.
- » The team interpreting the results.

The different types of blinding are summarized in **Table 28.3**.

Sometimes, an intervention cannot be masked. In such cases, the study team must make all attempts to **minimize potential sources of bias**.

Table 28.3. Blinding Types in Clinical Trials

Blinding Type	Characteristics
Unblinded or open label	All parties are aware of the treatment the participant receives
Single-blinded	Only the participant is unaware of the treatment he is receiving
Double-blinded	The participant and the clinicians/data collectors are unaware of the treatment the participant is receiving
Triple-blinded	Participant, clinicians/data collectors and outcome adjudicators/data analysts are all unaware of the treatment the participant is receiving

Randomization

Randomization is the process of **assigning** participants to treatment and control groups, assuming that **each participant has an equal chance of being assigned to any group.**

This idea was introduced by Fisher in a 1926 agricultural study. Since then, the academic community has deemed randomization an **essential tool for unbiased comparisons** of treatment groups.

Why Randomize?

1. Members in the groups should not differ in any systematic way.
 - In a clinical trial, if treatment groups are systematically different, trial results will be biased.
2. Proper randomization ensures no a priori knowledge of group assignment (i.e., allocation concealment).
 - That is, researchers, participants, and others should not know to which group the participant will be assigned.
 - Knowledge of group assignment creates a layer of potential **selection bias** that may taint the data.
3. Random assignment is necessary and guarantees **validity** for statistical tests of significance used to compare treatments.

How is Randomization Achieved?

Several randomization techniques have been proposed for the random assignment of participants to treatment groups in clinical trials, such as **simple randomization, block randomization, stratified randomization,** and **covariate adaptive randomization.** Again, their description is outside the scope of this book.

Properties of Randomization

The first property of randomization is that it **promotes comparability among the study groups**.

- » Such comparability can only be attempted in observational studies by adjusting for or matching on known **covariates**, with no guarantee or assurance of control for other covariates.
- » Randomization, however, extends a high probability of comparability with respect to **unknown important covariates** as well.

The second property is that the act of randomization provides a **probabilistic basis for an inference** from the observed results when considered in reference to all possible results.

Early Termination of a RCT

Sometimes, it is useful to include a rule to finish the study earlier than expected when the result is already **clear enough**. In these situations, it is **unethical** for a group of subjects to continue receiving treatment that has been shown to be **less effective** or **more harmful**.

These types of rules are usually incorporated in most studies with a high number of patients and which entail a follow-up of several years.

To achieve this, the results of the study must be **monitored**, while intermediate analyzes are carried out at predetermined times to consider whether the continuation of the study is likely to produce more conclusive or comprehensive responses.

Key Terms

Define the following terms.

Adaptative design	External validity	Phase IV studies
Another active treatment	Factorial design	Placebo
Bias	Group-randomized trial design	Post-marketing surveillance
Blinding	Informed consent	Randomization
Comparison group	Internal validity	Randomized Clinical Trial (RCT)
Control group	Intervention	Selection criteria
Crossover design	Masking	Sequential trial design and interim analysis
Double dummy	Non-inferiority study	Stopping rule
Early termination of a RCT	Parallel groups design	Superiority study
Efficacy	Phase I studies	Uncertainty principle
Enriched enrollment design	Phase II studies	Usual care
Equipoise	Phase III studies	
Equivalence study		

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. State the features of a RCT study design.
2. List some important advantages to using the RCT study design.
3. List some important disadvantages to using the RCT study design.
4. List the estimates that can be obtained with a RCT study.
5. Draw a diagram of the RCT study design.
6. What are the primary benefits of randomization?
 - Conducted to determine the safety of a treatment in humans. Patients go through intense monitoring.
 - Large studies (may or may not be a randomized trial) conducted after the therapy has been approved by the FDA to assess the rate of serious side effects and explore further therapeutic uses.
 - Relatively large randomized blinded trials used to evaluate the efficacy of an intervention.
 - Investigator explores test tolerability, safe dosage, side effects, and how the body copes with the drug.
7. Match the following definitions with (a) Preclinical, (b) Phase I, (c) Phase II, (d) Phase III, or (e) Phase IV.
 - Studies involving animals or cell cultures.

8. Experimental studies can involve therapeutic or preventive trials.

- Provide an example for each of these types of trials.

9. What are the primary benefits of blinding?

10. Multiple choice questions.

1. Randomized clinical trials (RCTs):

- a) Are not required to be based on the concept of equipoise.
- b) Always have a control arm that uses placebo.
- c) Are considered to be the gold standard for determining efficacy and safety in clinical research.
- d) Are always double blinded.

2. What do you mean by a randomized design?

- a) The subjects do not know which study treatment they receive.
- b) Patients injected with placebo and active doses.
- c) Randomly assigning subjects either for placebo or active dose.
- d) Signed document of the recruited patient for the clinical trial procedures.

3. What is meant by a blind subject?

- a) The subjects do not know which study treatment they receive.
- b) Patients injected with placebo and active doses.
- c) Fake treatment.
- d) Signed document of the recruited patient for the clinical trial procedures.

4. What is a “double dummy”?

- a) The subjects do not know which study treatment they receive.
- b) Patients injected with placebo and active doses.
- c) Fake treatment.
- d) Signed document of the recruited patient for

the clinical trial procedures.

5. What is placebo?

- a) The subjects do not know which study treatment they receive.
- b) Patients injected with placebo and active doses.
- c) Fake treatment.
- d) Signed document of the recruited patient for the clinical trial procedures.

6. What is informed consent in a clinical trial?

- a) The subjects do not know which study treatment they receive.
- b) Patients injected with placebo and active doses.
- c) Fake treatment.
- d) Signed document of the recruited patient for the clinical trial procedures.

7. Which one of the following is the last step of a clinical trial process?

- a) Investigator selection.
- b) Patient recruitment.
- c) Statistical Analysis.
- d) Data filed and registration.

8. How many people will be selected for phase I trial?

- a) The whole market will be under surveillance.
- b) 300-3000 people.
- c) 20-300 people.
- d) 20-50 people.

9. How many people will be selected for phase II trial?

- a) The whole market will be under surveillance.
- b) 300-3000 people.
- c) 20-300 people.
- d) 20-50 people.

10. How many people will be selected for phase III trial?

- a) The whole market will be under surveillance.
- b) 300-3000 people.
- c) 20-300 people.
- d) 20-50 people.

11. Which one of the following will be checked under phase IV surveillance?

- a) The whole market will be under surveillance.
- b) 300-3000 people.
- c) 20-300 people.
- d) 20-50 people.

12. What is “blinding” and what is its purpose?

- a) Blinding means you begin with the null hypothesis, and base your conclusions totally on a statistical analysis of the data without any preconceived ideas.
- b) Blinding refers to equipoise, i.e. uncertainty regarding whether a new treatment is effective.
- c) Blinding means that the subjects and/or investigators do not know which treatment group the subject is in. The purpose is to prevent bias in assessing the outcome.
- d) Blinding occurs when the results totally disagree with previously published studies. Its purpose is to cause a re-evaluation of the data.

13. What are the TWO main purposes of randomization?

- a) To eliminate bias in assignment to a treatment group.
- b) To achieve baseline comparability between the intervention and comparison (control) groups, i.e. to make the groups being compared similar with respect to known and unknown confounders.
- c) To avoid the problem of random error.
- d) To enhance the predictive value of the study.

14. In randomized trials of new human immunodeficiency virus therapies, investigators may use a composite endpoint known as the time to loss of virological response. Patients are deemed to meet the endpoint after the first of a series of events occurs: a new acquired immunodeficiency syndrome event, death, the patient is lost to followup or the patient experiences virological failure on treatment. At that point, the patient exits the trial and followup ceases on the patient. Which one of the following statements is true?

- a) Investigators use a composite endpoint as they cannot make a decision in advance about which is the most important outcome.
- b) Composite endpoints simplify the analysis of randomized trials.
- c) If one or more components of the composite endpoint are deemed to have greater clinical relevance than others, then appropriate analytical methods which take this into consideration must be used when analyzing a trial that utilizes a composite endpoint.
- d) A study that uses such a composite endpoint can provide reliable information about the frequency of occurrence of each component of the composite; thus, this type of trial provides good value for money.
- e) If a composite endpoint is used instead of basing the analysis on each component of the composite, the length of the trial must be increased.

15. Which one of the following statements about evidence-based medicine is true?

- a) The hierarchy of evidence indicates that a randomized controlled trial always provides stronger evidence than a cohort study.
- b) Published papers always provide all the relevant information (e.g. on diagnosis, prognosis or therapy) required for an evidence based investigation.
- c) Using evidence-based medicine means that a novel therapy will not be adopted in the community unless a relevant randomized controlled trial shows a statistically significant effect when compared to a control therapy.
- d) The number needed to treat (or harm) expresses the effectiveness (or safety) of an intervention in a way that is clinically meaningful.
- e) An evidence-based approach is restricted to conventional therapeutic interventions and can never be applied to alternative medicine.

16. Why is sample size important in clinical trials?

- a) With a large sample size, the effects of confounding are minimized.
- b) An adequate number is needed to show a clinically important treatment effect.
- c) Large clinical trials reduce bias and therefore provide more reliable results.
- d) Results from trials with a large sample size are more ethical.
- e) Larger numbers of participants tend to increase treatment differences between groups.

17. Which of the following statements about blinding is true:

- a) It helps prevent measurement bias - the biased assessment of outcomes.
- b) It helps reduce selection bias.
- c) It is required to do an ITT analysis.
- d) It is required for concealment.

18. In a randomized controlled trial, the following types of biases are reduced by randomization:

- a) Ascertainment bias.
- b) Selection bias.
- c) Recall bias.
- d) Publication bias.
- e) Bias in handling dropouts.

19. Researchers plan to evaluate a new oral immune-modulating therapy for locally advanced breast cancer. They conceive a RCT to test the new treatment in women who have hormone receptor-negative, human epidermal growth factor receptor 2 (HER2)-negative cancers. Currently accepted treatment for this condition includes radiotherapy and intravenous chemotherapy. Which of the following comparison groups would best preserve blinding in a randomized trial of the new agent while maintaining equipoise?

- a) No treatment.
- b) A placebo.
- c) Radiotherapy and intravenous chemotherapy plus a placebo.
- d) Radiotherapy and intravenous chemotherapy.
- e) Delayed treatment.

20. Excessive exclusion criteria may result in which of the following:

- a) Decreased external validity.
- b) Increased internal validity.
- c) Problems with recruitment.
- d) Increased costs.
- e) All of the above.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. *N Engl J Med* 2016; 375:65-74.
- Brody T. Clinical Trial Design. In: Brody T. Clinical trials. Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. 2ª Edition. Cambridge: ELSEVIER; 2016.
- Evans SR. Fundamentals of clinical trial design. *Exp Stroke Transl Med*. 2010 Jan 1; 3(1): 19–27.
- Hróbjartsson A, Emanuelsson F, Thomsen ÅSS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *Int J Epidemiol*. 2014 Aug;43(4):1272-83.
- Kang M, Ragan BG, Park JH. Issues in Outcomes Research: An Overview of Randomization Techniques for Clinical Trials. *Journal of Athletic Training*. 2008;43(2):215–221.
- Karanickolas PJ, Farrokhyar F, Bhandari M. Blinding: Who, what, when, why, how? *Can J Surg*. 2010 Oct; 53(5): 345–348.
- Patino CM, Ferreira JC. Internal and external validity: can you apply research study results to your patients? *J Bras Pneumol*. 2018 May-Jun; 44(3): 183.
- Rosenberg WF, Lachin JM. Randomization in Clinical Trials. Theory and Practice. 2ª Edition. New Jersey: Wiley; 2016.
- Stoney CM, Johnson LL. Design of Clinical Trials and Studies. In: Gallin JI, Ognibene FP, Johnson LL. Principles and Practice of Clinical Research. 4ª Edition. United Kingdom: ELSEVIER; 2018.
- Umscheid CA, Margolis DJ, Grossman CE. Key Concepts of Clinical Trials: A Narrative Review. *Postgrad Med*. 2011 Sep; 123(5): 194–204.

Equivalence, and Non-inferiority Trials

Learning objectives for this chapter

- A. Distinguish between superiority, non-inferiority and equivalence trials.
- B. Differentiate the methodology between superiority, non-inferiority and equivalence trials.
- C. Identify the biostatistics behind the non-inferiority and equivalence trials.
- D. Define and state the importance of the non-inferiority margin.
- E. Apply the non-inferiority and equivalence trials into clinical practice.

As it is fundamentally **impossible** to demonstrate that two treatments are equal, since the 1970s new methodological procedures have come up, allowing the development of evidence studies destined to show the **absence of significant differences** between treatments, and non-inferiority trials.

Sometimes, a Randomized Clinical Trial (RCT) is not designed with the objective of determining the superiority of the intervention under study in relation to a comparison intervention, but rather to determine if they are the same (**equivalence studies**) or if, at least, the first is not inferior to the second (**non-inferiority studies**).

The essential difference between superiority, equivalence and non-inferiority trials relies in the **formulation of the hypothesis to be tested**. **Table 29.1** depicts the algorithms of analysis for these three types of trials. If **T** represents the measure of efficacy of the new treatment, and **C** stands for the efficacy of the control treatment (more on efficacy in the following chapter), therefore:

- » **Superiority trials:** T is superior to C.
- » **Non-inferiority trials:** The difference between C and T is smaller than a margin “**M**”.
- » **Equivalence trials:** The difference between C and T is not smaller or bigger than a margin “**M**”.

Basically, the word “equivalent” means not inferior and not superior.

Table 29.1. Hypothesis Formulation for Superiority, Non-inferiority, and Equivalence Trials

Study Type	Null Hypothesis (H ₀)	Alternate Hypothesis (H ₁)
Superiority	$C - T \geq 0$	$C - T < 0$
Non-inferiority	$C - T \geq M$	$C - T < M$
Equivalence	$ C - T \geq M$	$ C - T < M$

Pharma Bridge to



Bioequivalence studies, correspond to phase I trials. They are carried out by the pharmaceutical industry to compare two formulations or methods of administration of a drug, with the intention of demonstrating that they are interchangeable. The response variables used are pharmacokinetic measures. They are usually carried out with a small number of subjects and using crossed designs.

In 1995, Altman and Bland stated: "Absence of evidence is not evidence of absence".

Equivalence Trials

Most RCTs aim to determine whether one intervention is **superior** to another. Failure to do so **does not implies they are equivalent**.

Equivalence trials aim to determine whether one (typically new) intervention is **therapeutic similar** to another (usually an existing) treatment with respect to predefined clinical criteria.

Lesaffre defined **equivalence** as a difference in performance of two therapies for which "the patient will not detect any change in effect when replacing one drug by the other."

The design of equivalence trials is **similar**, but not entirely the same, to that of **bioequivalence trials**.

Biostatistics Behind an Equivalence Trial

Equivalence implies that the **new mean** is only **slightly better or worse** than the old mean.

Performing a standard t-test and finding that it does not disprove the null hypothesis **is not a substitute for equivalence testing** because it may merely reflect a low power.

If **both t-tests** are compatible with the null hypothesis, then the observed difference lies **within the permissible difference**, so that the two drugs have **equivalent effects**.

A variant of this test is used to calculate **confidence limits** for the difference between the two means. If this lies **within** the limits $\pm\Delta$, which demarcates a zone of scientific or clinical indifference, **equivalence is demonstrated**. This principle is shown in **Figure 29.2**.

Hypothesis Testing in Equivalence Trials

Suppose a group of clinicians agree that two therapies are equivalent if the observable difference (ΔE) between them lies within an established interval defining clinical equivalence (from $-\Delta$ to $+\Delta$).

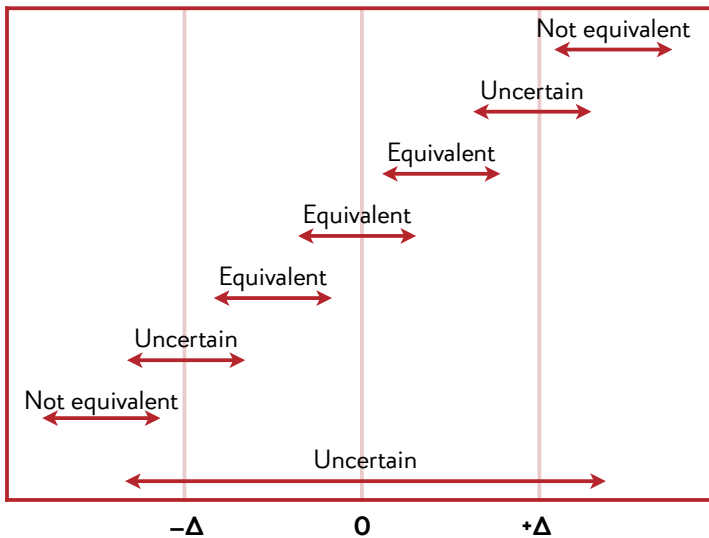


Figure 29.1. Range from $-\Delta$ to $+\Delta$ within which equivalence is assumed.

Data on the true difference (ΔE) between the therapies are collected and analyzed to **reject** the null hypothesis ($H_0: \Delta E > +\Delta$ or $\Delta E < -\Delta$), favoring the alternative hypothesis by means of a suitable statistical test, such as the **chi-square test** or **Student's t-test**.

Clinical Example of an Equivalence Trial

Robertson, et al. conducted an open-label, pragmatic equivalence trial to compare Dihydrocodeine and Methadone for opioid maintenance treatment. Patients were recruited from multiple locations such as the Edinburgh and Lothian Community Drug Problem Service and the practice sites of two general practitioners in the United Kingdom. Opioid dependency was defined as the detection of at least one positive test result during urinary toxicology analysis. Patients recommended for opioid maintenance treatment were recruited for the study. However, patients who were pregnant or had psychiatric complications were excluded from the trial. The patients were randomized to receive either a Methadone mixture (1 mg/mL) or Dihydrocodeine (30- or 60-mg tablets). The primary endpoint was **retention** in treatment.

Power calculations estimated that a sample size of 250 patients would yield 80% power for testing the **hypothesis** that the Dihydrocodeine group would retain 10% fewer patients than the Methadone group and produce a 0.5-point increase on the opioid dependency scale.

A total of 235 patients were available for the first treatment. In the first follow-up session, over 90% of the study participants were available for consultations. In the final follow-up phase (36 months after initial treatment assignment), information was obtained from 84% of the study participants.

Statistical analyses revealed that the 95% CI of the difference in retention proportion between the randomized groups included zero and that the treatments did not differ significantly from each other after 6 months (−5.2%, 12.6%), 12 months (−10.8%, 10.3%), or 18 months (−1.1%, 25.0%). Repeated measurement analysis indicated that the 95% CI of the difference in the score assessing illicit opioid usage (averaged over all follow-up periods) was −0.31 to 0.29, also indicating that the treatments did not differ significantly for this variable.

Non-inferiority Trials

As we've stated at the beginning of this chapter, it is not statistically possible to prove that two treatments are identical. Nevertheless, it is possible to determine that a new treatment **is not worse** than a reference treatment by more than an acceptable amount, and with a given degree of confidence.

Non-inferiority of a new treatment with respect to the reference treatment is of interest on the premise that the new treatment has some other **advantages**, such as:

- » Greater availability.
- » Reduced cost.
- » Less invasiveness.
- » Fewer adverse effects (harms).
- » Greater ease of administration.

The new treatment will be recommended if it is **similar** to the reference treatment for a prespecified primary outcome, but not if it is **worse** by more than Δ .

Biostatistics Behind a Non-Inferiority Trial

The **null hypothesis** in a non-inferiority study states that the primary end point for the experimental treatment **is worse** than that for the positive control treatment by a prespecified margin.

In order to demonstrate non-inferiority, the null hypothesis must be **rejected at a pre-specified level of statistical significance**, in favor of the alternate hypothesis with an adapted classical statistical test.

Some non-inferiority trials have been criticized for merely studying a new marketable product ("me-too" drugs) without offering any advantages over existing products.

The **H₀ for non-inferiority** is that the new treatment is inferior to the reference treatment due to a difference higher than or equal to a pre-specified margin.

The **alternative hypothesis** in this type of studies is that the difference between treatments is **smaller** than the pre-specified margin.

Figure 29.2 outlines the statistical evaluation to be used and the range of possible outcomes for a trial designed to demonstrate non-inferiority.

If the confidence interval for the study results **excludes** the prespecified margin (i.e., the non-inferiority margin, also called “delta”), then the conclusion is made that the test treatment is **non-inferior** to the control.

- » Traditionally, the confidence interval is a **97.5% one-sided** or **95% two-sided** interval.
- » For simplicity, all the confidence intervals are considered **two-sided**.

The treatment is considered **non-inferior** if the lower limit of the 95%CI of the difference between treatment and control **does not include** the value of the pre-specified margin.

Features of Non-inferiority Studies

The major components of non-inferiority study designs are summarized in **Table 29.2**.

Non-inferiority Margin

The margin M quantifies the maximal loss of clinically acceptable efficacy for the studied treatment to be considered as non-inferior to the control.

- » **High** M values increase the probability that inferior treatments be considered as non-inferior.
- » **Lower** and conservative M values demand bigger samples, thus making the studies more expensive, and with ethical implications.

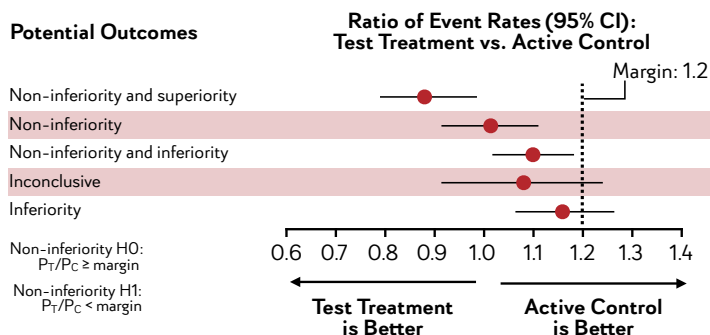


Figure 29.2. Hypothesis Testing in Non-inferiority Trials.

Table 29.2. Features of Non-inferiority Studies

Consideration	Explanation	Challenges
Active control	Select active control on the basis of a previous randomized superiority trial comparing active control with placebo; active control represents current standard of care	Placebo-controlled trials may not have been performed
End-point selection	Is the end point clinically relevant, and are there historical data comparing the active control with placebo for the selected end point?	Composite end points may be difficult to interpret; the relevance of end points may change in the course of follow-up
Choice of non-inferiority margin	Is the margin less than the treatment effect of the active control versus placebo? Is there consensus about the margin of reduced effectiveness that is still acceptable in light of potential benefits (e.g., improved safety, lower cost, lower risk of side effects)?	It is important not to accept new therapies that are less effective over time than previous therapies (known as "biocreep"); historical data are not always available to determine the difference between placebo and control (e.g., in the case of anti-infective agents)
Assay sensitivity	If the active control were compared with placebo, would superiority be evident?	A "positive control" usually cannot be assessed in the study, since placebo is not feasible or ethical
Constancy and metrics	Have the conditions changed between the trial establishing superiority of the active control over placebo and the non-inferiority trial? What type of metric (between-group difference in absolute risk or relative risk) is more likely to be constant between studies and therefore a reliable metric for comparison and margin definition?	Characteristics of the study population or concomitant therapies may have changed since the effect of active therapy was established, making a determination of non-inferiority unreliable; constancy is not always present for absolute effects; a lower-than-expected event rate may make a risk-difference margin clinically inappropriate if viewed from a relative-risk perspective; a higher-than-expected event rate may result in lower- than-expected power
Execution	Are the assigned treatments administered adequately? Is ascertainment of the end point accurate and complete?	Lack of attention to execution in the control group or misclassification or missing data on the end point may bias the study toward a conclusion of non-inferiority
Analysis	If treatment crossover or non-adherence occurs, what is the appropriate analysis (intention-to-treat or per-protocol)?	Treatment crossover may bias an intention-to-treat analysis toward a conclusion of non-inferiority, but a per-protocol analysis may also introduce bias, since baseline characteristics are no longer balanced between study groups

The M value must be established based on **clinical and statistical considerations**, and must be defined **prior** to the study.

In a simple manner, M may be determined as a percentage of the control effect estimated for the current study, usually **between 10 and 20%**. However, its definition must take into consideration the therapeutic field and the magnitude of the control group effect.

» For **anti-infectious** agents, more conservative margins are recommended (e.g. 10%) when the expected effect is around 90%, and more ample margins (e.g. 20%) when the anticipated effect is inferior to 80%.

The non-inferiority margin may also be determined by the so-called “**50%-rule**”, endorsed by the Food and Drug Administration (FDA), which advocates that the value of M must be **inferior** (preferentially 50%) **to the lower limit** of the 95%CI obtained from historical data that compare control treatment and placebo.

Clinical Example of a Non-Inferiority Trial

In patients with atrial fibrillation, Warfarin reduces the risk of stroke, as compared with placebo or Aspirin, but is associated with an increased risk of bleeding and requires frequent blood testing to ensure a therapeutic effect. Several new oral anticoagulant agents are associated with a lower risk of bleeding and offer greater convenience, since they do not require blood testing. These agents have recently been examined and approved by the FDA on the basis of three large non-inferiority trials comparing the oral anticoagulants with Warfarin for the prevention of stroke or thromboembolism: ARISTOTLE (Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation), RE-LY (Randomized Evaluation of Long-Term Anticoagulant Therapy), and ROCKET-AF (Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation).

Prior randomized trials of Warfarin versus Aspirin provided the expected rate of stroke or systemic thromboembolism. The non-inferiority trials compared new anticoagulants with Warfarin in study populations ranging from 14,264 to 18,261 participants randomly assigned to treatment groups, with the relative risk of stroke or thromboembolism as the primary end point and a relative non-inferiority margin of less than 1.4. The upper bounds of the one-sided 97.5% confidence interval for the relative risk in each study ranged from 0.95 to 1.11, falling below the pre-specified margin and supporting the conclusion of non-inferiority in each trial. These studies also showed less frequent intracranial hemorrhage, which, along with greater convenience for patients, has led to the replacement of warfarin with these new anticoagulants as first-line therapy to prevent stroke in many patients with atrial fibrillation.

Key Terms

Define the following terms.

Confidence limits

Equivalence

Equivalence trials

Non-inferiority margin

Non-inferiority trials

Superiority trials

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. State the features of an equivalence trial.
2. State the features of a non-inferiority trial.
3. Based on Table 29.1, explain with your own words the hypothesis testing for superiority, non-inferiority, and equivalence studies.
4. What is a “me-too” drug? Search the web for some examples.
5. What would be the advantages that could lead a research team to perform a non-inferiority trial?
6. Search the web for an example of a non-inferiority and an equivalence trial.

Bibliography and Suggested Reading

- Argimon-Pallás JM. Métodos de investigación clínica y epidemiológica. 4ª edición. Barcelona: ELSEVIER; 2013.
- Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. *J Hepatol.* 2007; 46:947–54.
- Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2009;361:1139–51.
- Drazen JM, Harrington DP, McNyrat JVV, Ware JH, Woodcock J. Challenges in the Design and Interpretation of Noninferiority Trials. *N Engl J Med* 2017;377:1357–67.
- Fleming TR. Design and interpretation of equivalence trials. *Am Heart J.* 2000; 139:S171–6.
- Granger CB, Alexander JH, McMurray JVV, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2011;365:981–92.
- Hahn S. Understanding noninferiority trials. *Korean J Pediatr.* 2012 Nov; 55(11): 403–407.
- Greene WL, Concato J and Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? *Ann Intern Med.* 2000; 132:715–22.
- Hoffman JIE. t-Test Variants: Cross-Over Tests, Equivalence Tests. In: Hoffman JIE. *Basic Biostatistics for Medical and Biomedical Practitioners.* Waltham: Elsevier Academic Press; 2019.
- Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis.* 2008;66(2):150–4.
- Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med* 2011; 365:883–91.
- Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. Extension of the CONSORT 2010 statement. *JAMA.* 2012; 308(24): 2594–2604.
- Pinto VF. Non-inferiority clinical trials: concepts and issues. *J Vasc Bras.* 2010;9(3):145–151.
- Robertson JR, Raab GM, Bruce M, et al. Addressing the efficacy of dihydrocodeine versus methadone as an alternative maintenance treatment for opiate dependence: a randomized controlled trial. *Addiction.* 2006; 101:1752–9.
- Sedgwick P. Equivalence trials. *BMJ* 2013;346:f184.

Efficacy, Effectiveness, Efficiency

Learning objectives for this chapter

- A. Explain how the outcomes of a health intervention can be measured.
- B. Explain the importance of measuring efficacy, effectiveness, and efficiency.
- C. Define efficacy, effectiveness, and efficiency.
- D. Understand efficacy, effectiveness, and efficiency studies as a way to improve quality in health interventions and organizations.
- E. Understand the intention-to-treat (ITT) principle and determine its usefulness in clinical trials.
- F. Understand the per-protocol (PP) principle and determine its usefulness in clinical trials.
- G. Apply efficacy, effectiveness, and efficiency studies into clinical practice.

Public health systems aim to improve the health of a population at a cost whose limits are set, one way or another, by society. To ensure that the health of a population is improved, clinical procedures that are supported by scientific evidence of a more favorable outcome (risk/benefit ratio) compared to any other alternative should be used.

As we have learned in the previous chapters, the best scientific evidence on the outcome of therapeutic interventions comes from well organized **randomized clinical trials** (RCTs).

Besides efficacy, effectiveness and efficiency, the **outcome** of a health intervention can be measured in:

- » **Utility:** Survival time adjusted for quality of life.
- » **Benefit:** Outcomes expressed in monetary units.

Archie Cochrane, the British pioneer and famous clinical epidemiologist, defined the three basic concepts related to testing healthcare interventions: Efficacy, effectiveness, and efficiency.

Efficacy asks the question: "Can it work?"

Study Validity
Internal validity: The degree to which the results are attributable to the independent variable and not some other rival explanation.

External validity: The extent to which the results of a study can be generalized.



Efficacy

Provides evidence carried out under specific circumstances or “**under ideal conditions**”.

Efficacy is obtained by a canonical methodology in a **RCT**.

Is necessary for effectiveness, but it alone is not sufficient.

Efficacy Studies

Efficacy studies provide a “**truer**” effect of the intervention itself. This serves as a proof-of-concept or testing whether the program has the potential to be impactful or how big the size of that impact could be.

They are also known as **Phase III studies**, where the intervention under study is **compared against a placebo**.

These studies have a **high internal validity**.

» **Key question** in internal validity: can the observed changes be attributed to your program or intervention (i.e., the cause) and not to other possible causes?

On the other hand, efficacy studies may have **lower generalizability** or **external validity**. That is, if the study is replicated there are low probabilities that the same results will be obtained.

Example

Imagine you are studying how a financial incentive program improves smoking cessation rates. Your research question can be: will paying people to quit smoking and stay abstinent improve long-term quit rates?

You decide to perform this study as a RCT. You set the inclusion and exclusion criteria, and after the recruitment time, homogenize the patient sample and randomize it to financial incentives or not (as you can see, this is a **highly controlled situation**).

At the end of the study, you realize a 10% higher quit rates in the group with the program. Now, the most important thing to determine is that the financial incentive were **actually the cause** of that higher quit rate.

Randomization and sample selection were used to increase the internal validity of the study. However, you are worried that, when the program is reproduced, other researchers may not get the same results.

Assume a scenario where other researchers do not obtain the same results as you. How did this happen?

We are experiencing a crisis of replicability in science. We find it difficult to replicate some major influential study findings. This is in part related to the methodology of the studies (efficacy vs effectiveness), but there are many other factors that play too: publication bias, mining data for p values, etc.

Maybe, the participants in their study were younger employees who haven't been smoking for the same length of time than those individuals in your trial. Or maybe the replicated study had different environmental factors than yours. This decreases the **external validity** of the study.

Effectiveness

Provides evidence from interventions that take place in normal circumstances or in “**real life**”.

Depends not just on efficacy, but also on local factors which may differ from those of the clinical trial (technology, experience, organization, etc).

Effectiveness asks the question:
“Does it work in practice?”

Effectiveness Studies

Effectiveness studies provide estimates of the program's impact in “real life”. That is, estimates that might be **highly relevant to replicate**. However, there might be other factors at play that may explain the observed effect (the effect could be **confounded**).

They compare the intervention under study **against another currently used intervention**.

This studies have a **high external validity**.

Example

Imagine you are assessing the impact of using bundled payments to pay for joint replacement surgery on cost and quality. In a Medicare program, hospitals volunteered to be paid a fixed price for the acute hospital care plus any services a patient used over the 90 days that followed. This was implemented at hundreds of hospitals across a country. The participants had to contend with what post-acute care providers were available to them, their capacity, whether there were other bundled payers in the market, etc.

A drug that has efficacy is efficient when the patient uses it based on the established regimen.

That being said, you may worry there is something totally separate driving the effect you are observing about bundled payments improving cost and quality. What if the providers were also Accountable Care Organizations, explaining why they were doing so well? Remember the participants volunteered to be in the program. What if there is a selection bias that could affect the causal estimate of the impact? This makes us question the **internal validity** of the study.

Efficiency asks the question:
"Is it worth it?"

Efficiency

Measures the effect of an intervention in relation to the **resources** it consumes.

It is the ratio of the outcomes to costs that have to be met to achieve the outcome.

Efficiency Studies

Efficiency studies are more often called **cost-effectiveness** or **cost-benefit studies**, and are carried out by the Economists. Therefore, their description is outside the scope of this book.

Shifting the Relative Strengths of the Studies

What is worse?

- » To **claim a treatment effect which actually does not exist**, and thus, to potentially jeopardize patients with an inefficacious therapy, or
- » To conclude that the **efficacy of an actual efficacious therapy cannot be proven** and, as a consequence, to potentially refuse patients an efficacious therapy.

From a patient's perspective, the answer might not be so straightforward. However, there are **two principles** that will help us obtain a clear answer to this question in clinical research.

Intention-to-Treat (ITT) Principle

The ITT principle states that **every randomized patient** in the clinical study **should enter the primary analysis**.

Accordingly, patients who drop out prematurely, who are non-compliant to the study treatment, or that even take the wrong study treatment, are included in the primary analysis within the respective treatment group they have been assigned to at randomization.

Consequently, analyzing data according to the ITT principle, the original randomization and the number of patients in the treatment groups remain **unchanged**, the analysis population is **as complete as possible**, and a potential **bias** due to exclusion of patients is **avoided**.

However, there are only two specific reasons that might cause an **exclusion of a patient** from the full analysis set:

- » **No treatment** was applied at all.
- » There are **no data** available after randomization

Per-Protocol (PP) Principle

While an analysis according to the ITT principle aims to preserve the original randomization and to avoid potential bias due to exclusion of patients, the aim of a PP analysis is to identify a **treatment effect which would occur under optimal conditions**. Therefore, some patients from the full analysis set need to be excluded from the population used for the PP analysis (named **PP population**).

Usually, this applies to patients fulfilling any of the following criteria:

- » Any major **protocol deviations** (e.g. intake of a concomitant medication affecting the primary endpoint).
- » Non-availability of measurements of the **primary endpoint**.
- » Non-sufficient exposure to **study treatment**.

Both approaches, the ITT and the PP approach, are **valid**, but have different roles in the analysis of clinical studies.

Let's come back to the question at the beginning of this section:

What is worse, **scenario A** (claim a non-existing effect) or **scenario B** (neglect an existing effect)?

To answer this, consider the **essential difference** between the two scenarios:

- » **Scenario A** means that a **statistically proven result is actually wrong** (a result that might cause dangerous effects).
 - Based on such a proof, an inefficacious treatment might be approved and patients put into danger.
- » **Scenario B** means that **efficacy was not proven but also not refused**.
 - However, the non-proven efficacy does not equal a proven inefficacy.
 - From a scientific perspective, such a non-decision has less implications than a wrong proof.



Bridge to Statistical power

Type I error: Rejecting the null hypothesis when it is in fact true.

Type II error: Not rejecting the null hypothesis when it is in fact not true.

Concluding, it is more essential to **avoid a wrong proof** than to avoid a wrong non-decision (which is also bad, but scenario A is worse). Consequently, it is essential to keep the probability of scenario A below the level of significance (e.g. 5%).

Key Terms

Define the following terms.

Benefit	Efficiency	principle
Cost-effectiveness	External validity	Internal validity
Effectiveness	Generalizability	Per-protocol (PP) principle
Efficacy	Intention-to-treat (ITT)	Utility

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Draw a table containing the terms “efficacy”, “effectiveness”, and “efficiency”, their definition and the study design from which they are obtained.
2. Draw a table comparing internal validity vs. external validity.
3. Get together with some classmates and discuss in what ways does efficiency trials play a foundational role in public health?
4. Get together with some classmates and discuss pros and cons of using an intention-to-treat analysis.
5. Get together with some classmates and discuss pros and cons of using a per-protocol analysis.
6. Search the web for articles where the researchers used the ITT or the PP analysis.

Bibliography and Suggested Reading

- Briel M, Montori VM, Dureux P, Devereaux PJ, Guyatt G. The Principle of Intention to Treat and Ambiguous Dropouts. In: Guyatt G, Rennie D, Meade MO, Cook DJ. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. 3rd Edition. New York: Mc Graw-Hill Education; 2015.
- Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. *Fundamentals of Clinical Trials*. 5th Edition. New York: Springer; 2015.
- Kim SY. Efficacy versus Effectiveness. *Korean J Fam Med*. 2013 Jul; 34(4): 227.
- McCoy CE. Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *West J Emerg Med*. 2017 Oct; 18(6): 1075–1078.
- Piantadosi S. *Clinical Trials. A Methodologic Perspective*. 3rd Edition. New Jersey: John Wiley & Sons; 2017.
- Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspect Clin Res*. 2016 Jul-Sep; 7(3): 144–146.
- Singal AG, Higgins PDR, Waljee AK. A Primer on Effectiveness and Efficacy Trials. *Clin Transl Gastroenterol*. 2014 Jan; 5(1): e45.
- Sedgwick P. What is per protocol analysis? *BMJ* 2013;346:f3748.

Bias

Learning objectives for this chapter

- A. Define error and its types (random and systematic error).
- B. Distinguish between random error and bias.
- C. Distinguish between imprecision and inaccuracy.
- D. Define bias.
- E. Understand how bias affect the results of a clinical study.
- F. Identify the possible sources of bias in clinical research.
- G. Identify the different types of bias.
- H. Understand selection bias.
- I. Understand information bias.
- J. Learn to assess bias in clinical studies.

While the results of an epidemiological study may reflect the true effect of an exposure(s) on the development of the outcome under investigation, it should always be considered that the findings may in fact be **because of an alternative explanation**.

Those alternative explanations may be due to the effects of:

- » **Chance** (random error).
- » **Bias**.
- » **Confounders** (discussed in **Chapter 20**).

Whatever the explanation, **spurious results** are produced, leading to conclude the existence of a **valid statistical association** when it does not exist, or alternatively the absence of an association when one is truly present.

Observational studies are the **most susceptible** to the effects of these alternative explanations, therefore they must be considered at both the design and analysis of the study in order to minimize their effects.

Error

Error is defined as the **difference between the true value of a measurement and the recorded value of a measurement**. There are **many sources** of error in collecting clinical data.

Error can be described as **random** or **systematic**.

Random Error

Is also known as **variability**, **random variation**, or “**noise in the system**”.

The heterogeneity in the human population leads to relatively large random variation in clinical trials.

Random error has **no preferred direction**, so we expect that averaging over a large number of observations will yield a net effect of zero. The estimate may be **imprecise, but not inaccurate**.

The impact of random error, **imprecision**, can be minimized with **large sample sizes**.

Random error corresponds to **imprecision** in **Figure 31.1**.

Systematic Error

Is also known as “**bias**”.

Refers to **differences that are not attributable to chance alone**.

» A simple example: Picture a measuring device that is improperly calibrated, so that it consistently overestimates (or underestimates) the measurements by X units.

Systematic error has a **net direction and magnitude**, so averaging over a large number of observations does not eliminate its effect.

Bias can be large enough to invalidate any conclusions, and increasing the sample size will not help to minimize it.

In human studies, bias can be **subtle and difficult to detect**. Even the suspicion of bias can render judgment that a study is invalid. Thus, the design of clinical trials focuses on removing known biases.

Bias corresponds to **inaccuracy** in **Figure 31.1**.

Precision:

Refers to how close a bunch of replicate measurements come to each other (that is, how reproducible they are).

Accuracy:

Refers to how close your measurement tends to come to the true value, without being systematically biased in one direction or another.



Figure 31.1. Graphic difference between accuracy and precision.

Biases

Bias literally means **distortion of statistical result**, but for the purpose of epidemiological studies it has been defined as **deviation of results, or inferences from the truth, or process leading to such deviation**.

Biases can lead to under-estimation or over-estimation of the true intervention effect and can vary in magnitude: some are **small and trivial** (compared with the observed effect), and some are **substantial** (so that an apparent finding may be due entirely to bias). This converges in **lacking the internal validity** in the study, and a study is valid if its results correspond to the truth.

Stages of Research in Which Bias can Occur

There are many sources of bias, but there are **seven major scenarios** that could originate bias in clinical research:

1. In **reading-up** on the field.
2. In **specifying** and selecting the study sample.
3. In **executing** the experimental manoeuvre (or exposure).
4. In **measuring** exposures and outcomes.
5. In **analyzing** the data.
6. In **interpreting** the analysis.
7. In **publishing** the results [and back to 1].

Another definition for **Bias** could be:
 "Any systematic error in design, conduct or analysis of study that results in mistaken estimate of an exposure's effect on risk of disease".

Bridge to Study Validity

Internal validity:

Can I rely on the conclusions of this study?

External validity:

Can I apply these conclusions to my patients?

Validity is an expression of the degree to which a test is capable of measuring what it intends to measure.

You can access the "Catalogue of Bias" from the University of Oxford by clicking the following link: <https://catalogofbias.org/biases/>

Types of Bias

In 1979, Sackett summarized a draft of a catalog of **35 biases**. By April 28th 2020, the University of Oxford, on the other hand, listed **58 different biases** in its website "Catalogue of Bias". This has led the reader to a difficult approach to bias over the years, because of the long number of biases, and the absence of a formal consensus regarding their definitions. That's why one of the purposes of this chapter is to provide a clear (and brief) theoretical framework of biases and some practical tips that will help you assess bias in clinical studies.

The most common types of bias are summarized in **Table 31.1**, but we will deepen in two major categories of bias: **selection bias** and **information bias**.

Selection Bias

Also known as **Berksonian bias**.

Participants in research may differ systematically from the population of interest. For example, participants included in an influenza vaccine trial may be healthy young adults, whereas those who are most likely to receive the intervention in practice may be elderly and have many comorbidities, therefore they are not representative of the target population.

Similarly, in observational studies, conclusions from the research population may not apply to "real-world" people, as the observed effect may be exaggerated or it is not possible to assume an effect in those not included in the study.

Selection bias can arise in studies because groups of participants may **differ in ways other than the interventions or exposures under investigation**. When this is the case, the results of the study are biased by **confounding (Chapter 20)**.

Selection bias can have **varying effects**, and the magnitude of its impact and the direction of the effect is often hard to determine.

Assessing Selection Bias

To **assess selection bias**, authors should **include the following** information at different stages of the trial or study:

- » Numbers of participants screened as well as randomized/ included.
- » How intervention/exposure groups compared at baseline.
- » To what extent potential participants were re-screened.
- » Exactly what procedures were put in place to prevent prediction of future allocations and knowledge of previous allocations.

Table 31.1. Classification of Bias in Clinical Trials with Examples and Remedies

Type of Bias	Example	Remedies
Selection bias	Favoring the assigning of patients known by the investigator to the treatment group	Appropriate randomization
Study management or performance bias	Following more closely the patients in the treatment group favored by the investigator	Blinding, when feasible Standardization of procedures Personnel training
Detection bias	Recording outcomes in a way that proves the investigator's or the participant's beliefs	Blinding, when feasible
Attrition or loss-to-follow-up bias	Participant loss related to the outcome (e.g., severe side effects)	Intention-to-treat analysis
Publication or reporting bias	Selective reporting of only statistically significant results	Trial registration, prepublication trial protocol, reporting also negative results and not only positive results

- » What the restrictions were on randomization (e.g. block sizes).
- » Any evidence of unblinding.
- » How missing data from participants lost to follow-up were handled.

Remember that **randomization** of participants in experimental studies aims to provide the fairest method of comparing the effect of an intervention with a control, and preventing selection biases is part of this aim. However, sometimes it may not be perfectly achieved.

Because anything that happens after randomization can affect the chance that a study participant has the outcome of interest, it is essential that all patients (even those who fail to take their medicine or accidentally or intentionally receive the wrong treatment) are analyzed in the groups to which they were allocated (**intention-to-treat principle**; **Chapter 30**).



Intention-to-treat principle: Assessment of the people taking part in a trial, based on the group they were initially (and randomly) allocated to, regardless of whether or not they dropped out, fully adhered to the treatment or switched to an alternative treatment.

Types of Selection Bias

Some authors consider **three types** of selection bias:

- » **Incidence-prevalence bias (Neyman bias):** A late look at those exposed (or affected) early will miss fatal and other episodes, plus mild or silent cases and cases in which evidence of exposure disappears with disease onset.
 - **Example:** You are studying an association between diabetes mellitus and renal failure. Cases will be interviewed one month after occurrence of renal failure, but renal failure patients with diabetes die more frequently. The remaining cases of renal failure would show lower frequency of diabetes mellitus, thus under-evaluating the association between renal failure and diabetes.
- » **Loss-to-follow-up bias:** Occurs in prospective cohort studies when individuals lost to follow-up do not have the same probability of having the clinical outcome of interest in comparison with individuals who remain under observation.
 - **Example:** You are performing a prospective study aimed at determining the incidence rate of renal insufficiency in hypertensive and normotensive patients. Because the follow-up duration must be extended to several years, normotensive patients who did not develop any disease after several years of observation may be less stimulated to continue the study. On the other hand, hypertensive patients, who most likely develop comorbid conditions, can be more motivated to continue the study participation.
- » **Publication bias:** Occurs most commonly in systematic reviews and meta-analysis. It occurs due to the influence of the study results on chances of publication. Studies with positive results are more likely to be published than studies with negative results leading to a preponderance of false positive results in the literature.

Information Bias

Information bias is any **systematic difference from the truth** that arises in the collection, recall, recording and handling of information in a study, including how missing data is dealt with.

All types of study can be subject to information bias, but **observational studies**, particularly in those with retrospective designs, are at greater risk because rely on retrospective data collection.

Assessing Information Bias

To **assess information bias**, an appropriate study design must be chosen, following well-designed protocols for data collection and handling, and the appropriate definition of exposures and outcomes.

Experimental designs are not excluded of this type of bias. Therefore, ensuring that **blinding** of intervention status is maintained whilst outcomes are measured and recorded is a key element to minimize information bias (**Chapter 28**).

Types of Information Bias

There are **four major types** of information bias:

- » **Misclassification bias:** Any systematic difference from the truth arising from recording participants or features of interest in the wrong category. This can lead to an underestimation of the prevalence, as well as of potential risk factors.
 - **Example:** You are performing a screening study. Patients who are at low risk for the condition are not screened, and classified as negative for the condition.
- » **Observer bias:** Systematic difference between a true value and the value actually observed due to observer variation.
 - **Example:** In the assessment of medical images, one observer might record an abnormality but another might not.
- » **Recall bias:** Occurs when participants do not remember previous events or experiences accurately or omit details: the accuracy and volume of memories may be influenced by subsequent events and experiences.
 - **Example:** Parents of children diagnosed with cancer may be more likely to recall infections earlier in the child's life than parents of children without cancer.
- » **Reporting bias:** This term covers a range of different types of biases, and has been described as the most significant form of scientific misconduct. A general definition could be the distortion of presented information from research due to the selective disclosure or withholding of information by parties involved with regards to the topic selected for study and the design, conduct, analysis, or dissemination of study methods, findings or both.
 - Furthermore, researchers have described **seven types of reporting bias** (publication bias, time-lag bias, duplicate publication bias, location bias, citation bias, language bias, and outcome reporting bias), whose discussion is outside the scope of this book.

– **Example:** In 2015, Jones et. al, compared the outcomes of randomized controlled trials specified in registered protocols with those in subsequent peer-reviewed journal articles. There were discrepancies between pre-specified and reported outcomes in 30% of the studies, and 13% of trials introduced a new outcome in the published articles compared with those specified in the registered protocols.

Key Terms

Define the following terms.

Accuracy

Attrition bias

Bias

Chance

Detection bias

Error

Imprecision

Incidence-prevalence bias

Information bias

Inaccuracy

Loss-to-follow-up bias

Misclassification bias

Observer bias

Performance bias

Precision

Publication bias

Random error

Recall bias

Reporting bias

Selection bias

Systematic error

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

- 1. Answer:** What are the most susceptible studies to the effect of an alternative explanation?
- 2. Draw a table** comparing random error vs. systematic error.
- 3. List the seven major steps** that could originate bias in clinical research.
- 4. Access the “Catalogue of Bias”** from the University of Oxford and get familiar with all the types of bias.
- 5. State the differences** between precision and accuracy, and complete **Figure AL31.1** in order to fully understand those differences.



Figure AL31.1. Targets to graph the difference between accuracy and precision.

Bibliography and Suggested Reading

- Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
- Catalogue of Bias Collaboration, Nunan D, Bankhead C, Aronson JK. Selection bias. *Catalogue Of Bias* 2017: <http://www.catalogofbias.org/biases/selection-bias/>
- Catalogue of bias collaboration. Bankhead CR, Spencer EA, Nunan D. Information bias. In: Sackett *Catalogue Of Biases* 2019. <https://catalogofbias.org/biases/information-bias/>
- Delgado-Rodríguez M, Llorca J. Bias. *Journal of Epidemiology & Community Health* 2004;58:635-641.
- Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Med.* 2015; 13: 282.
- Kumar G, Acharya AS. Biases in epidemiological studies: How far are we from the truth? *Indian Journal of Medical Specialities.* 2014;5(1):29-35.
- Lambert J. Statistics in Brief: How to Assess Bias in Clinical Studies? *Clin Orthop Relat Res.* 2011 Jun; 469(6): 1794-1796.
- Pandis N. Sources of bias in clinical trials. *Am J Orthod Dentofacial Orthop.* 2011;140:595-6.
- Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; 32: 51-63.
- Sanderson S, Tatt ID, Higgins JPT. Tools for Assessing Quality and Susceptibility to Bias in Observational Studies in Epidemiology: A Systematic Review and Annotated Bibliography. *Int J Epidemiol.* 2007;36(3):666-76.
- Tripepi G, Jager KJ, Deijer FW, Wanner C, Zoccali C. Bias in clinical research. *Kidney International.* 2008;73:148-153.

Section VII

Evidence Synthesis

Chapters of the Section

Chapter 32	The Role of Evidence Synthesis in Health Care
Chapter 33	Systematic Reviews and Meta-Analysis
Chapter 34	Clinical Practice Guidelines
Chapter 35	Quality of Evidence

The Role of Evidence Synthesis in Health Care

Learning objectives for this chapter

- A. Define evidence synthesis.
- B. Understand the importance of evidence synthesis in the current healthcare practice.
- C. Describe the central role of evidence synthesis in evidence-based health care.
- D. Identify where evidence synthesis can be found.
- E. Identify the core of evidence synthesis used in clinical practice.

Health systems worldwide face increasingly **complex challenges**, such as the growing burden of chronic non-communicable diseases, climate change and the emergence of new epidemics and antimicrobial resistance. These challenges have prompted an important shift in focus **from curative care to prevention** and health promotion, as well as the development of new service delivery, financing and governance models.

Meeting these challenges will require new policies and health systems reforms that are informed by **robust and contextualized evidence**. This process will, in turn, rely upon the **synthesis** and **appraisal** of a wide array of research information and knowledge stemming from various data sources.

Evidence synthesis is a tool used by researchers and health care providers everyday and plays a really important role in making informed clinical decisions.

We often hear in the news or social media about exciting or controversial new findings from the latest study or trial. However, we hope you recognize by now that it is very important that we do not focus on one single study or piece of evidence in order to make a clinical decision. All the information you have gathered over the last chapters is intended for you to recognize that **one study on its own can be inaccurate or misleading**, and it may not always give the full picture. That's why using evidence synthesis is a way of making sure that your **views and clinical decisions** are based on findings of **lots of studies** rather than just one.

In short, evidence synthesis has an important role to play in informing how health care decisions are made.

One of the multiple benefits that evidence synthesis has made to healthcare is the use of steroids given to women who are about to give birth prematurely, giving the newborn the chance of survival.

There are **many types** of evidence synthesis, however they all share the same aims to provide users with clean, clear, trustworthy, and useful information. Some of the most important include:

- » **Narrative reviews.**
- » **Systematic reviews.**
- » **Meta-analysis.**
- » **Clinical practice guidelines.**

Nevertheless, **not all evidence synthesis are equal**, and we need to discriminate between a good one from a bad one. This will be discussed in a following chapter.

The Core of Evidence Synthesis

Research evidence is generated through **primary studies** that typically use either quantitative, qualitative methods or, in some instances, a combination of the two. The process of gathering together evidence generated through primary research studies is referred to as **systematic review**, and represents the **core** of evidence-based practice worldwide (more on this in the following chapter).

The process of synthesizing research evidence is not limited to the dichotomy of quantitative and qualitative methods. Instead, a **wide spectrum of methods** is available to address one or more of the following aims:

- » Aggregate information.
- » Explain or interpret processes, perceptions, beliefs and values.
- » Develop theories.
- » Identify gaps in the literature or the need for future research.
- » Explore methodological aspects of a method or topic.
- » Develop or describe frameworks, guidelines, models, measures or scales.

Where is Evidence Synthesis?

There are several organizations dedicated to support and collate evidence synthesis relevant to public health policy and practice. These are summarized in **Table 32.1**.

Table 32.1. Different Organizations that Support Evidence Synthesis

Organization	Description
The Cochrane Collaboration	This global organization consists of around 40 review groups, each with a focus on a different health topic. Cochrane reviews are usually focused on the effectiveness of clinical interventions, and take a somewhat narrow approach to evidence synthesis.
The Campbell Collaboration	Has a similar approach to the Cochrane Library, but focuses more on education, social welfare and development.
The EPPI-Centre	Is part of the Institute of Education at the University College London (UCL). This unit does systematic reviews for different government departments, with a wide variety of topics and methods. The Centre has also developed novel synthesis techniques, software, and review methods including participatory methods.
3ie	Funds, produces, quality assures and synthesises rigorous evidence on development effectiveness. They support evaluations and reviews that examine what works, for whom, why, and at what cost in low- and middle-income countries.
Centre for Reviews and Dissemination (CRD)	Established at York University. Performs health-relevant systematic reviews and has developed particular expertise in high quality systematic reviews and associated economic evaluations.
The Health Evidence Network (HEN)	Started by the WHO/Europe in 2003, produces a variety of publications to meet policy-makers' needs for evidence and synthesizing the best available evidence in response to policy-makers' questions. These include joint policy briefs and policy summaries, produced with the European Observatory on Health Systems and Policies, which synthesize the evidence around specific policy options for tackling key health system issues; and HEN summaries of reports, including synopses of the main findings and policy options.

Key Terms

Define the following terms.

Evidence synthesis

Location of evidence synthesis

Primary studies

Systematic review

Uses of evidence synthesis

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Answer: What are the main types of evidence synthesis?

2. Answer: What are the main aims available to address with evidence synthesis?

3. Browse through the different organizations where you can find evidence synthesis and get familiar with them.

Bibliography and Suggested Reading

- Carvalho G. Introduction of the Evidence synthesis: article type. *Proc. R. Soc. B.* 2018;285: 20180858.
- Chandler J, Cumpston M, Thomas J, Higgins JPT, Deeks JJ, Clarke MJ. Chapter I: Introduction. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated August 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
- Langlois EV, Daniels K, All EA. *Evidence Synthesis for Health Policy and Systems: A Methods Guide*. Geneva: World Health Organization; 2018.
- Pearson A. Evidence synthesis and its role in evidence-based health care. *Nurs Clin North Am.* 2014 Dec;49(4):453-60.
- Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol.* 2009; 9, 29.

Systematic Reviews and Meta-Analysis

Learning objectives for this chapter

- A. Understand the importance of systematic reviews in clinical research.
- B. Understand the levels of evidence of clinical research.
- C. Distinguish between a narrative review and a systematic review.
- D. Distinguish the types of primary studies that must be included in the different types of systematic reviews.
- E. Identify the steps involved in developing the systematic review protocol.
- F. Understand the different types of bias associated with a systematic review.

The modern world and its globalization process have generated a growing and constant appearance of new information, reflected in **multiple articles and publications**. This reality has also involved biomedical sciences, which have observed an increase in the number of articles that accredit the use of therapies and treatments endorsing their uses.

Given the large number of articles and publications available, the simplest and most complete way of using this information is by **compiling it**.

A **systematic review of scientific evidence** consists of the **synthesis of the best available evidence**, which purpose is to address a precisely defined research question using all available studies in a specific field (usually referred to as **primary studies**), and by applying an explicit and rigorous methodology.

Systematic reviews use formal, structured, and unbiased methods for synthesizing scientific evidence. Unlike narrative reviews, systematic reviews are characterized by **transparency**, **reproducibility**, and the **capability of being repeatedly updated**.

A systematic review is a study of studies.

Narrative reviews often contain major flaws such as selective use of evidence and subjective criteria for drawing conclusions.

The three fundamental characteristics of a Systematic Review are:

- Transparency
- Reproducibility
- Capability of being updated.

Many systematic reviews contain meta-analyses, but not all.

Although the terms “**systematic review**” and “**meta-analysis**” are often used interchangeably, their precise definition is mandatory:

» **Systematic review** applies to the **entire research process of the investigation study**.

– A systematic review attempts to **collate all empirical evidence** that fits pre-specified eligibility criteria to **answer a specific research question**.

– It uses explicit, systematic methods that are selected with a view to **minimizing bias**, thus providing reliable findings from which conclusions can be drawn and decisions can be made.

» **Meta-analysis** is the use of **statistical techniques** to integrate and **summarize the results of included studies** and obtain a **joint estimate**.

Evidence Pyramid

Systematic reviews and meta-analysis are situated at the **top** of what is known as the “**Evidence Pyramid**” (Figure 33.1).

» As you move up the pyramid the amount of available literature on a given topic **decreases**, but the relevancy and quality of that literature **increases**.

Applications of Systematic Reviews and Meta-Analyses

Systematic reviews and meta-analyses are essential tools for summarizing evidence accurately and reliably. Their **uses** include:

- » Help clinicians keep up-to-date.
- » Provide evidence for policy makers to judge risks, benefits, and harms of health care behaviors and interventions.
- » Gather together and summarize related research for patients and their carers.
- » Provide a starting point for clinical practice guideline developers.
- » Provide summaries of previous research for funders wishing to support new research.
- » Help editors judge the merits of publishing reports of new studies.

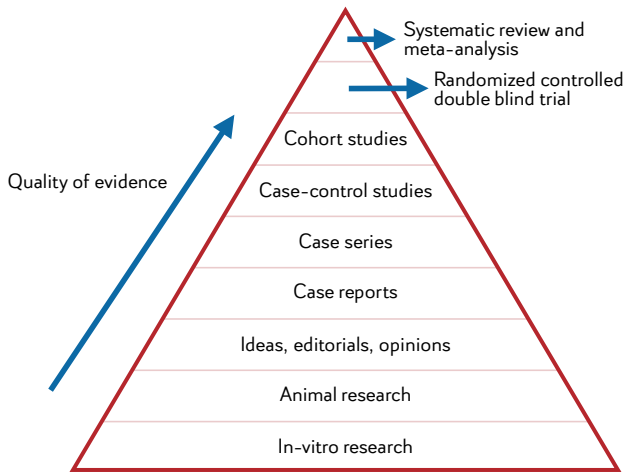


Figure 33.1. The Evidence Pyramid.

Study Design for Systematic Reviews

The **design** of the primary studies included in a systematic review is **determined by the research question** that wants to be answered:

- » **Effect of therapy/intervention:** Randomized Controlled Trials (RCTs).
- » **Causes of disease (etiology):** Cohort or case-control studies.
- » **Diagnosis:** Cross-sectional studies.
- » **Prognosis:** Cohort studies.
- » **Prevalence:** Cross-sectional studies.
- » **Experience of disease:** Qualitative studies.

Stages of a Systematic Review

A systematic review is a research study that, unlike those presented in previous chapters of this book, is **not based on primary data**. Rather, it **uses data previously collected** in other studies. Therefore, they can be considered as **observational** studies in which the “study population” is made up of the best original articles on the subject under review.

As with any other study designs, the planning of a systematic review requires the elaboration of a **protocol** that details the definitions and procedures that will be carried out throughout its different stages.

You can access the PRISMA website by clicking the following link: <http://prisma-statement.org>

Study Validity

Internal validity:

Is the extent to which the study answers its research question.

External validity:

Refers to the generalize the results of a study from the source population to the target population and is a measure of the practical utility of the results.



The **Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)** statement is the prevailing guideline for conducting and reporting systematic reviews. PRISMA focuses on ways in which authors can ensure the transparent and complete reporting of systematic reviews and meta-analyses. It does not address directly or in a detailed manner the conduct of systematic reviews, for which other guides are available. Also, PRISMA recommends that every systematic review protocol should be registered in a repository (or published in a suitable journal) prior to performing the review.

The systematic review protocol should include the items summarized in **Table 33.1**, and the published systematic review report itself should indicate where the systematic review protocol was registered and how to access it.

Validity of the Included Studies in a Systematic Review

If you want to draw appropriate conclusions from a systematic review, the included studies must be **valid**. Remember that validity can be divided into **internal** and **external validity**.

As the most important assessment for a systematic review is **bias**, not quality, it is no longer recommended that systematic reviews include traditional tools to assess study quality, such as scales or checklists. Instead, an **assessment of the risk of bias** should be performed. One way to assess bias is the **Cochrane Risk of Bias Tool**, which consists of six general domains (**Table 33.2**).

Effect Measure in Systematic Reviews

An **effect measure** is a statistical measure used to compare the outcomes of two treatment or exposure groups. It is used in **meta-analysis** to summarize the results at the study level and to calculate the **common effect** across the studies (**Table 33.3**).

It is useful to distinguish between **relative** (ratio-based) and **absolute** (difference-based) **effect measures**.

Relative effect measures, such as RR and OR, often suggest more optimistic treatment effects than do absolute effect measures, such as RD.

» **Example:** If 2 of 1000 patients in the control group and 1 of 1000 patients in the treatment group experience an unwanted event, the $RR = 0.5$ represents a risk reduction of 50%, whereas the absolute value of the $RD = 0.001$, an absolute risk reduction of 0.1%. The former arguably sounds more impressive than the latter. Whether or not this is a clinically interesting effect depends on the specific research problem,

Table 33.1. Phases for Conducting a Systematic Review**Definition of the objective (research question).****Search for evidence:**

- Databases consulted.
- Keywords used.
- Coverage time.
- Other requirements: journals, language, etc.

Inclusion and exclusion criteria of the studies.**Determination of the quality of the studies:**

- Summary of the essential characteristics of the studies.
- Evaluation of the quality of the studies.
- Assessment of variability among researchers who determine the quality of the studies.

Data collection

- Registration of the characteristics of the studies:
 - Type of article and year of publication.
 - Study design.
 - Characteristics of the intervention.
 - Characteristics of the control group.
 - Sample size.
- Recording of study results.

Analysis of the results:

- Homogeneity tests.
- Statistical combination of results.
- Tests to detect systematic differences between studies.
- Graphic representations.
- Sensitivity analysis.
- Subgroup analysis.

Conclusions and recommendations.

and factors such as patient preference, other outcomes, costs, and practical considerations.

For **binary outcomes**, the study design has a bearing on the choice of effect measure. Case-control studies can only estimate ORs, because the incidence of disease, which is needed to calculate risk, is unknown in case-control studies.

The natural effect measures for RCTs and cohort studies are RD and RR, respectively, although ORs are commonly used for both designs.

Table 33.2. The Six General Domains of the Cochrane Risk of Bias Tool

Domain	Explanation
Random sequence generation	Is the method used to generate the allocation sequence described in sufficient detail to allow an assessment of whether it should produce comparable groups?
Allocation concealment	Is the method used to conceal the allocation sequence described in sufficient detail to determine whether intervention allocations could have been foreseen in advance of or during enrollment?
Blinding of participants and personnel	Are the measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received described in sufficient detail to determine whether the intended blinding was effective?
Blinding of outcome assessment	Are the measures used, if any, to blind outcome assessors from knowledge of which intervention a participant received described in sufficient detail to determine whether the intended blinding was effective?
Incomplete outcome data	Are the participants included in the analysis exactly those who were randomized into the trial? Is the completeness of outcome data, including attrition and exclusions from the analysis, for each main outcome described? Are the reasons for attrition/exclusions reported?
Selective outcome reporting	Are there indications that the study authors have failed to report outcome data that seem sure to have been recorded?
Other potential threats to validity	Are there important concerns about bias not addressed in the other domains?

Table 33.3. Common Effect Measures for Meta-analyses

Binary outcomes	Risk difference (RD), relative risk (RR), and odds ratio (OR)
Continuous outcomes	Mean difference (MD), standardized mean difference (SMD), and mean ratio (MR)
Survival outcomes	Hazard ratio (HR), and median survival time (MST)
Incidence rates	Incidence rate ratio (IRR), and incidence rate difference (IRD)

Important Issues in Effect Measure

There are several other important issues related to the choice of effect measure in meta-analyses:

- » Ratio-based effect measures (e.g., RR, OR, IRR) often indicate a **more optimistic treatment effect** than do difference-based effect measures (e.g., RD, IRD).
- » Ratio-based effect measures have **greater stability across different risk** groups than do difference-based effect measures.
- » Difference-based effect measures reflect the **baseline risk** of individuals, whereas ratio-based effect measures do not.
- » If some of the included studies have zero events, **difference-based measures are better** than ratio-based measures.
- » For binary outcomes, OR has **better statistical properties** than RD and RR; however, OR is not as easy to interpret or communicate.
- » For continuous outcomes, SMD has the **best statistical properties**, MD is **most easily interpreted**, and MR is **preferable with skewed data**.

Bias in Systematic Reviews

Biases may appear in different phases of a systematic review, mainly in the **location and study selection**.

As discussed in **Chapter 31**, biases may **threaten the validity** of the conclusions.

The most important biases in systematic reviews are summarized in **Table 33.4**.

Table 33.4. Most Important Biases in Systematic Reviews

Bias	Explanation
Publication bias	<p>Defined as failure to publish the results of a study "on the basis of the direction or strength of the study findings." That is, not all studies have the same probability of being published.</p> <p>Frequently, authors decide not to send their manuscripts to the journals, or the editors and reviewers of a journal decide not to accept some studies based on certain characteristics, related more to the results found than to quality aspects.</p> <p>The prevention of this bias is important from two perspectives:</p> <ul style="list-style-type: none"> • The scientific one: to achieve a complete dissemination of knowledge. • The authors perspective: if articles with positive results are preferably published, any RS will tend to get positive results too.
Language bias	<p>The English language has been the predominant language in medical research.</p> <p>Publication in other languages can sometimes be regarded as of secondary importance.</p> <p>Studies publishing positive results might also be more likely to publish in English.</p> <p>Reading and using only English language research could provide a biased assessment of a topic, and can lead to biased results in systematic reviews.</p>
Database bias	<p>The two most widely used bibliographic databases, MEDLINE and EMBASE, do not have the same coverage and, therefore, if the search is limited to articles indexed only in one of them, a bias can be introduced.</p>
Citation bias	<p>In order to locate the studies that should be included in a systematic review, it is common for the authors to complement the search in the databases by contacting experts in the field and making a manual search from the bibliographic references of the published studies.</p> <p>Citation bias is likely to be introduced when performing this manual search, as studies with positive results are generally cited more frequently than studies with negative results.</p>
Multiple publication bias	<p>Studies with statistically significant results tend to be published more frequently; therefore, it is easier to locate and include them in a review. On the other hand, if they are not identified as multiple publication, duplicate data may be included leading to an overestimation of the effect.</p>

Key Terms

Define the following terms.

Absolute effect measure

Bias

Effect measure

External validity

Evidence pyramid

Internal validity

Meta-analysis

Narrative review

**Preferred Reporting Items
for Systematic Reviews and
Meta-Analyses (PRISMA)**

Primary studies

Relative effect measure

Systematic review

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Draw a table comparing systematic reviews vs. narrative reviews, and be sure to include definition, goals, question, components, number of authors, timeline, requirements, and values.
2. Take **Table 33.1** and make some cognitive bridges with the corresponding topic in this book, so that you are able to consolidate this knowledge.
3. With your classmates, find a scenario in which each of the biases inherent in systematic reviews can be applied.
4. State the three fundamental characteristics of systematic reviews.
5. Multiple-choice questions.
 1. Why do you need to review the existing literature?
 - a) You enjoy reading the academic research on your topic.
 - b) Because without it, you could never reach the required word-count.
 - c) To find out what is already known about your area of interest.
 - d) To make sure you have a long list of references.
 2. A systematic literature review is:
 - a) One which generates a literature review using a treasure hunt system.
 - b) A replicable, scientific, and transparent process.
 - c) One which gives equal attention to the principal contributors to the area.
 - d) A manufactured system for generating literature reviews tailored to your subject.
 3. Which of the follow is a benefit of a systematic review?
 - a) It reduces researcher bias and demands the researcher is comprehensive of their approach.
 - b) It is really quick to complete.
 - c) It is cost effective as an approach.
 - d) It provides internal validity to the study.

4. What is a limitation of systematic review?

- a) It is too hard to do.
- b) The research cannot be defined into the impact of a single variable.
- c) They are particularly complicated.
- d) The researcher community finds them complex.

5. What is distinctive about a narrative literature review?

- a) It is a review based exclusively on stories about companies, in book and case-study form.
- b) It is an historically-based review, starting with the earliest contributions to the field.
- c) It is a paraphrase style of reviewing which does not require referencing.
- d) It serves as a means of gaining an initial impression of a topic, which you will understand more fully as you conduct your research.

Bibliography and Suggested Reading

- Ahn E, Kang H. Introduction to systematic review and meta-analysis. *Korean J Anesthesiol.* 2018 Apr; 71(2):103–112.
- Argimon-Pallás JM. *Métodos de investigación clínica y epidemiológica.* 4ª edición. Barcelona: ELSEVIER; 2013.
- Balk E, Bonis PAL. Systematic review and meta-analysis. Elmore JG, ed. UpToDate. Waltham, MA: UpToDate Inc. <https://www.uptodate.com> (Accessed May 3rd, 2020).
- Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
- Brennan E, Cooper JC, Davis A. Evidence-Based Practice: Use in Answering Queries and developing Systematic Reviews. In: Thomas D. *Clinical Pharmacy Education, Practice and Research.* Clinical Pharmacy, Drug Information, Pharmacovigilance, Pharmacoeconomics and Clinical Research. Amsterdam: ELSEVIER; 2019.
- Fagerland MW. Evidence-Based Medicine and Systematic Reviews. In: Laake P, Benestad HB, Olsen BR. *Research in Medical and Biological Sciences.* From Planning and Preparation to Grant Application and Publication. Waltham: Elsevier Academic Press; 2015.
- Garg AX, Hackman D, Tonelli M. Systematic Review and Meta-analysis: When One Study Is Just not Enough. *CJASN.* 2008; 3(1) 253–260.
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med.* 2009;6(7): e1000100.
- Moreno B, Muñoz M, Cuellar J, Domancic S, Villanueva J. Revisiones Sistemáticas: definición y nociones básicas. *Rev. Clin. Periodoncia Implantol. Rehabil. Oral.* 2018;11(3): 184–186.
- Uman LS. Systematic Reviews and Meta-Analyses. *J Can Acad Child Adolesc Psychiatry.* 2011 Feb; 20(1): 57–59.

Clinical Practice Guidelines

Learning objectives for this chapter

- A. Learn about the origins of clinical practice guidelines.
- B. Understand the importance of clinical practice guidelines in healthcare.
- C. Identify the main sources of clinical practice guidelines and their regulations in Mexico.
- D. Identify the main sources of clinical practice guidelines worldwide.
- E. Identify the steps that comprises the clinical practice guideline methodology.
- F. Understand the attitudes and the acceptance of clinical practice guidelines in clinical practice.

Health professionals face a high number of problems in their daily activities, and most of the times they come simultaneously.

For example, suppose you have a patient with an herpetic lesion. Several questions can be asked:

- » Is it really herpetic?
- » Is it treated with Acyclovir?
- » At what dose the Acyclovir should be initiated?
- » Furthermore, knowledge of these problems may not be up-to-date or may not correspond to the best available evidence.

In clinical practice, it is usual to have to make multiple decisions in a short time, and under an enormous amount of pressure. Furthermore, there are also many issues where clinicians may have **different opinions** about the relative value of different treatment options or diagnostic strategies of a disease. In this daily occurrence, the **paradigm** is: “the decisions of health personnel by definition are correct”. However, when analyzing it, we find that these decisions may be **incorrect**, are **different**, and may be **not supported by the best available evidence**, and promotes the **variability** observed in clinical practice.

Summarizing, the **source of differences in the decision-making process** about individual patients can be found in one of the following:

- » **Uncertainty:** There is no quality scientific evidence on the value of possible treatments or diagnostic methods.
- » **Ignorance:** There is scientific evidence, but the clinician does not know or does not have it updated.
- » **External pressures:** The professional knows the value of tests or treatments, but uses other guidelines.
- » **Resources and services offered:** In the absence of the diagnostic technique or the recommended treatment, an alternative is used. The opposite is also true, because high availability can lead to overuse.
- » **Patient preferences:** in most cases the final decision is made by the patient or his family and his values and preferences also count, so the actions to perform may vary significantly from one patient to another.

It has been documented that a patient with a common condition receives **adequate care only half the time**, and that only **10 to 15%** of health interventions are **supported by adequate scientific research**, since every year two million articles are published in twenty thousand scientific journals, making it **difficult** for health personnel to select the information to keep updated.

Because of this high **variability in the clinical practice**, along with the struggle in managing and integrating the abundant information that frequently exists on a given medical problem, it seems logical that both the different health services and scientific societies, and those responsible for health policy, are concerned with having tools that allows the access to the **appropriate up-to-date information** in terms of quantity, and quality, in order to contribute to decision-making, seeking **equity** and **invariability** in healthcare.

It is in this setting that **clinical practice guidelines** (CPGs) made their appearance in the 1990s, with the intention to assist practitioner and patient decisions, **improve** the effectiveness of interventions and the quality of health care, and diminish variability in medical practice.

Clinical Practice Guidelines in Mexico

Starting in 2007, the **Programa Nacional de Guías de Práctica Clínica** emerged, under the coordination of the **Subsecretaría de Integración y Desarrollo del Sector Salud** through the **Centro Nacional de Excelencia Tecnológica en Salud (CENETEC)**.

Most of the questions in the Examen Nacional para Aspirantes a Residencias Médicas (ENARM) are based on CENETEC's CPGs.

This led the sectoral integration of the **Master Catalog of Clinical Practice Guidelines** as a national benchmark for promoting clinical and managerial decision-making based on recommendations established with the best scientific evidence available, aiming to reduce the variability of clinical practice, as well as the use of unnecessary and ineffective interventions, facilitate the maximum benefit in the treatment of patients and the lowest risk at an acceptable cost.

Clinical practice guidelines are updated on a scheduled basis from **3 years up to 5 years** after their publication in the Master Catalog of Clinical Practice Guidelines, or earlier, if there is new evidence that determines their renewal.

To achieve this end, the **Dirección de Integración de Guías de Práctica Clínica**, together with the public institutions of Health, creates:

- » Clinical Practice Guidelines.
- » Methodology for the Integration of Clinical Practice Guidelines.
- » Training Workshops in Evidence-Based Medicine.

Likewise, within the framework of improving the quality of the National Health System, and with the purpose of responding to the health needs of the rightful population, the **Instituto Mexicano del Seguro Social (IMSS)**, through the Dirección de Prestaciones Médicas, has considered as one of its strategic projects the **Development and Implementation of Clinical Practice Guidelines with a focus on Evidence-Based Medicine (EBM)**, in order to contribute to improving the quality of medical care, by giving greater importance to effective and safe interventions, based on scientific evidence.

As of May 8th 2020, the Master Catalog of Clinical Practice Guidelines by CENETEC has **343 guidelines**.

As of May 8th 2020, the project Development and Implementation of Clinical Practice Guidelines by IMSS has **523 guidelines**.

Clinical Practice Guidelines Worldwide

The last decade saw the origins of CPGs made and published by different academic, and governmental organizations and made them available to any user in repositories that, for the most part, do not need any registration or payment. **Table 34.1** summarizes sites where it is possible to access clinical practice guides from different **governmental and educational entities**.

Its use is similar in all of them: in the search window that is displayed when accessing the name of the pathology or clinical condition of interest is typed and the search engine will display the guides available in the repository.

Table 34.1. Organizations with Clinical Practice Guidelines Available

Organization	Country	Website
National Guideline Clearinghouse	United States of America	http://www.guideline.gov
National Institute for Health and Care Excellence (NICE)	United Kingdom	https://www.nice.org.uk/guidance
Scottish Intercollegiate Guidelines Network (SIGN)	Scotland	http://www.sign.ac.uk/our-guidelines.html
The Guidelines International Network (G-I-N)	Several countries	https://www.g-i-n.net/library
Australian National Health and Medical Research Council	Australia	https://www.nhmrc.gov.au/guidelines-publications

Overview of Clinical Practice Guidelines

The Institute of Medicine (IOM) defines CPGs as “statements that include recommendations, intended to optimize patient care, that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options”.

Based on this definition, CPGs consist of **two parts**:

- » **The foundation:** Is a **systematic review** of the research evidence bearing on a clinical question, focused on the strength of the evidence on which clinical decision-making for that condition is based.
- » **A set of recommendations:** Involving both the **evidence and value judgments** regarding benefits and harms of alternative care options, addressing how patients with that condition should be managed, everything else being equal.

Guidelines have largely focused on the **effectiveness of interventions**. Over time, however, they have paid more attention to the size of the effect and the balance between effects on the one hand and harms and costs on the other as well as on the feasibility of following guidelines.

It is essential that CPGs are **credible** by professionals, so they must be based on the **best scientific knowledge available** and be drawn up through an explicit procedure by expert panels with representation from all the groups involved. **Table 34.2** summarizes the main aspects to be taken into account with CPGs.

Table 34.2. Characteristics of a CPG that contributes to its use

1. Have a clear and defined aim.
2. Reach a previous consensus between all the participants, including their representatives.
3. Available evidence must be included in a clear and updated form.
4. CPG must be compatible with the regulations and values of the people to whom it is addressed.
5. Recommendations must be clear and precise.
6. The quality of the study must be assured (patient-based, with equity, accessible, effective, efficient, and secure).
7. There is a clear method for its update.
8. Allow flexible and adaptable use for individual patients.
9. Have an attractive structure and design.
10. Are easy to apply.

It is important that the development group of the CPGs take measures to **avoid biases, distortions or conflicts of interest**, as well as provide a clear explanation of the **relationship between the evidence**, the available options, health outcomes and the strength of the recommendations.

Developing Trustworthy Clinical Practice Guidelines

There is a sufficient international consensus on how to prepare evidence-based clinical practice guidelines. This **sequential process** begins with the **delimitation of the subject matter** of the guide, continues with the process of **formulating the recommendations** based on the synthesis and evaluation of the best research, and ends with consideration of the aspects of **editing, disseminating, and updating** the CPG. **Figure 34.1** summarizes this sequential process.

The design and elaboration of the methodology of a clinical practice guideline consists of **two complementary stages**:

1. Guide design planning.
2. The different phases that allow putting into practice the chosen design and preparing the guide.

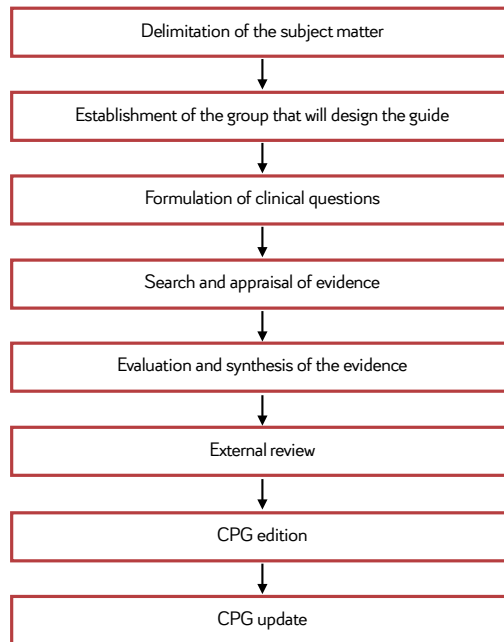


Figure 34.1. Sequential process to prepare an evidence-based CPG.

The team should be made up of professionals who will use the guide in their usual clinical practice and methodologists who will be in charge of analyzing the scientific evidence.

Planning

The composition of the **group** that will design the guide has to be multidisciplinary, both in terms of the number of professionals and the representation of disciplines. The group should have a coordinator and be **endorsed** by one or more professional societies and health institutions, which will facilitate the subsequent dissemination process of the guide.

CPGs on a certain subject are designed because there are a number of **reasons that justify the effort to carry it out**, such as:

- » Variability in clinical practice due to the existence of areas of uncertainty.
- » There is a major health problem with an impact on morbidity and mortality.
- » Emergence of novel techniques or treatments.
- » Possibility of achieving a change to improve results in care because:
 - The process is susceptible of being improved by a sanitary action.
 - The means to achieve this are available.

- » Availability of professionals with knowledge on the subject to be developed.
- » Priority area in the National Health Service.
- » Perceived need.

Conflicts of Interest

The CPGs present a synthesis of recommendations, and their content influences many health professionals. The readers of the CPGs must know the relationships that the individual authors of the guide maintain with the pharmaceutical industry as a way to guarantee **independence** and **transparency** in their development.

Before joining the working group, members and reviewers must make a statement about the existence or not of a **conflict of interest**.

A conflict of interest occurs in circumstances where **professional judgment** of a primary interest, such as patient safety or the validity of research, may be excessively **influenced** by secondary interests, whether this are financial benefit, desire to professional advancement, recognition of personal achievements and favors to friends, family, students or colleagues.

The usually declared interests are **financial**, not because they are more pernicious than others, but because they can be measured and objectively valued. The potential conflict of interest may exist regardless of whether the professional considers that these relationships have or does not influence his or her scientific judgment.

Other Aspects

The following points should be considered as part of the methodology for developing CPGs. However, their discussion is beyond the scope of this book.

- » Establish the scope and objectives of the CPG.
- » Specify tasks for team members.
- » Prepare a calendar and work plan (18-24 months on average).
- » Carry out a systematic review of the literature (based on what was studied in **Section II** and in **Chapter 33**).
- » Prepare the recommendations.
- » Writing, revision and update.

Attitudes and Acceptance of Clinical Practice Guidelines

Clinicians are **most likely** to accept recommendations from their own specialty society, **less likely** to trust those prepared by government agencies, and **least likely** to believe in guidelines prepared by managed care organizations and insurance companies.

Attitudes related to who prepared guidelines appear to be independent of the scientific validity of the guidelines themselves, although this has not been specifically studied.

Clinicians disagree on whether CPGs promote “**cookbook medicine**” with “**not enough recipes in the cookbook**” or evidence-based medicine. However, a majority also believe that guidelines are **biased, oversimplified, and rigid**, likely to decrease physician reimbursement, challenge physician autonomy, and decrease physician satisfaction.

Potential Benefits and Problems of Guidelines

Evidence-based, carefully developed, and updated guidelines provide many **potential benefits**, such as:

- » Synthesis of the literature by experts.
- » Clear recommendations for translating the evidence base into clinical application to foster best practice.
- » Opportunity to evaluate the outcomes of implementation in the “real world” setting.

However, several aspects of guidelines and their implementation need to be recognized as **potential problems**:

- » The challenge of keeping guidelines updated when the literature changes.
- » The potential for inappropriate use of guidelines for other than clinical purposes.
- » Difficulty accessing guidelines at the point of care – Many are lengthy or specific components relevant to a patient are not readily searchable or retrievable.
- » Lack of coordination among guideline development groups, generating differing recommendations.
- » Potential for conflicts of interest.
- » Application of guidelines developed to address a specific condition to patients with multiple comorbidities.

Quality Assessment of a Clinical Practice Guideline

The potential benefits of CPGs are only as good as the quality of the guidelines themselves. Appropriate methodologies and rigorous strategies in the guideline development process are important for the successful implementation of the resulting recommendations. The quality of guidelines can be extremely **variable** and some often fall short of basic standards.

The **Appraisal of Guidelines for REsearch & Evaluation (AGREE)** Instrument was developed in 2004 to address the issue of **variability in guideline quality**. The refined **AGREE II** instrument is a tool that assesses the **methodological rigour** and **transparency** in which CPGs are developed by local, regional, national or international groups or affiliated governmental organizations. These include original versions of and updates of existing guidelines.

The AGREE II consists of **23 key items** organized within **6 domains** followed by **2 global rating items** (“Overall Assessment”). Each domain captures a unique dimension of guideline quality.

- » **Domain 1. Scope and Purpose:** Concerned with the overall aim of the guideline, the specific health questions, and the target population (items 1-3).
- » **Domain 2. Stakeholder Involvement:** Focuses on the extent to which the guideline was developed by the appropriate stakeholders and represents the views of its intended users (items 4-6).
- » **Domain 3. Rigor of Development:** Relates to the process used to gather and synthesize the evidence, the methods to formulate the recommendations, and to update them (items 7-14).
- » **Domain 4. Clarity of Presentation:** Deals with the language, structure, and format of the guideline (items 15-17).
- » **Domain 5. Applicability:** Pertains to the likely barriers and facilitators to implementation, strategies to improve uptake, and resource implications of applying the guideline (items 18-21).
- » **Domain 6. Editorial Independence:** Is concerned with the formulation of recommendations not being unduly biased with competing interests (items 22-23).
- » **Overall assessment:** Includes the rating of the overall quality of the guideline and whether the guideline would be recommended for use in practice.

You can access the AGREE II Instrument and User's Manual by clicking the following link:
<https://www.agreetrust.org/resource-centre/>

Each of the AGREE II items and the two global rating items are rated on a **7-point scale** (1– strongly disagree to 7–strongly agree). A quality score is calculated for each of the six AGREE II domains.

The six domain scores are **independent** and should not be aggregated into a single quality score. The discussion and interpretation of domain scores with the AGREE II instrument is outside the scope of this book.

Key Terms

Define the following terms.

AGREE II

CENETEC

Clinical practice guidelines

Conflicts of interest

Planning a clinical practice guideline

Sources of difference in decision-making process

Variability in guideline quality

Variability in the clinical practice

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Enter the CENETEC webpage and become familiar with it. Find a CPG of your interest and identify the two editions that it includes: evidence and recommendations and quick reference.
2. Enter the IMSS CPG resource webpage and become familiar with it. Find a CPG of your interest and identify the two editions that it includes: evidence and recommendations and quick reference.
3. Remember a scenario based on your clinical practice where the variability in clinical practice, and in clinical decision-making, is clearly exemplified. Try to imagine a different outcome for that situation if clinical practice guidelines were used.
4. Remember a scenario based on your clinical practice where a conflict of interest arose. Try to imagine a way in which that conflict of interest could have been managed.
5. Enter the AGREE II website and download the User's Manual. Apply it to the CPG of your interest in Exercise 1 or 2 and determine its methodological quality.
6. List the characteristics of a CPG that contributes to its use in clinical practice.

Bibliography and Suggested Reading

- AGREE Next Steps Consortium (2017). The AGREE II Instrument [Electronic version]. Retrieved <05, 08, 2020> from <http://www.agreetrust.org>.
- Centro Nacional de Excelencia Tecnológica en Salud. Catálogo maestro de guías de práctica clínica. Available at: <http://cenetec-difusion.com/gpc-sns/>.
- Centro Nacional de Excelencia Tecnológica en Salud. Dirección de Integración de Guías de Práctica Clínica; 2017. Available at: <https://www.gob.mx/salud/%7Ccenetec/acciones-y-programas/direccion-de-integracion-de-guias-de-practica-clinica>.
- Eddy DM, Adler J, Patterson B, Lucas D, Smith KA, Morris M. Individualized guidelines: the potential for increasing quality and reducing costs. *Ann Intern Med*. 2011 May 3;154(9):627-34.
- Gisbert JP, Alonso-Coello P, Piqué JM. ¿Cómo localizar, elaborar, evaluar y utilizar guías de práctica clínica? *Gastroenterol Hepatol*. 2008;31(4):239-57.
- Gordillo-Moscoso A. Decision-Making in Medicine. Using Clinical Practice Guidelines. 2019. Resource of the Clinical Epidemiology Course. Facultad de Medicina de la Universidad Autónoma de San Luis Potosí.
- Grol R, Wensing M, Eccles M, Davis D. Improving patient care. The implementation of change in health care. 2nd Edition. Oxford: Wiley Blackwell; 2013.
- Jovell AJ, Navarro-Rubio MD, Aymerich M, Serra-Prat M. Metodología de diseño y elaboración de guías de práctica clínica en atención primaria. *Aten Primaria*. 1997;20(5):259-266.
- Berg AO, Atkins D, Tierney W. Clinical Practice Guidelines in Practice and Education. *J Gen Intern Med*. 1997 Apr; 12(Suppl 2): S25–S33.
- Peterson PN, Rumsfeld JS. The evolving story of guidelines and health care: does being NICE help? *Ann Intern Med*. 2011;155(4):269.
- Ransohoff DF, Pignone M, Sox HC. How to decide whether a clinical practice guideline is trustworthy. *JAMA*. 2013 Jan;309(2):139-40.
- Shekelle P. Overview of clinical practice guidelines. Aronson MD, ed. UpToDate. Waltham, MA: UpToDate Inc. <https://www.uptodate.com> (Accessed on May 8th, 2020).

Quality of Evidence

Learning objectives for this chapter

- A. Define quality of evidence and strength of recommendations.
- B. Understand the importance of determining the quality of evidence in clinical practice guidelines in healthcare.
- C. Learn about the GRADE approach for rating the quality of evidence.
- D. Learn about the Canadian Task Force on the Periodic Health Examination to establish the Levels of Evidence.
- E. Understand how quality of evidence can be transformed into recommendations.
- F. Learn about different ways to assess the strength of recommendations in clinical practice guidelines in healthcare.

Judgments about evidence and recommendations are complex, but a systematic and explicit approach to making judgments about the **quality of evidence** and the **strength of recommendations** can help to prevent errors, facilitate critical appraisal of these judgments, and can help to improve communication of this information.

Since the 1970s a growing number of organizations have employed various systems to grade the quality (level) of evidence and the strength of recommendations of clinical practice guidelines. Unfortunately, different organizations use **different systems** to grade the quality of evidence and the strength of recommendations.

The GRADE Approach

The GRADE approach is a system for rating the quality of a body of evidence in **systematic reviews** and **other evidence syntheses**. GRADE offers a transparent and structured process for developing and presenting evidence summaries and for carrying out the steps involved in developing recommendations. It is widely used to develop **clinical practice guidelines** and other **health care recommendations**.

Although the GRADE approach makes judgments about **quality of evidence**, that is confidence in the effect estimates, and **strength of recommendations** in a systematic and transparent manner, it **does not eliminate the need for judgments**. Thus, applying the GRADE approach does not minimize the importance of judgment or as suggesting that **quality can always be objectively determined**.

Although the quality of evidence represents a **continuum**, the GRADE approach results in an assessment of the quality of a body of evidence in **one of four grades** (Table 35.1).

The GRADE approach to rating the quality of evidence begins with the **study design** (trials or observational studies) and then addresses five reasons to possibly **rate down** the quality of evidence (Table 35.2) and three to possibly **rate up** the quality (Table 35.3).

Other Ways to Establish Levels of Evidence

Levels of evidence were originally described in a report by the **Canadian Task Force on the Periodic Health Examination** in 1979. The authors developed a system of rating evidence (Table 35.4) when determining the effectiveness of a particular intervention. The evidence was taken into account when grading recommendations.

Nonetheless, the levels of evidence were further described and expanded by **Sackett** in an article on levels of evidence for antithrombotic agents in 1989 (Table 35.5).

Quality of evidence is a **continuum**; any discrete categorization involves some degree of **arbitrariness**. Nevertheless, advantages of simplicity, transparency, and vividness outweigh these limitations.

Table 35.1. GRADE's grades for Quality of Evidence

Grade	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very Low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Table 35.2. Factors That can Reduce the Quality of the Evidence

Factor	Consequence
Limitations in study design or execution (risk of bias)	Decreases 1 or 2 levels
Inconsistency of results	Decreases 1 or 2 levels
Indirectness of evidence	Decreases 1 or 2 levels
Imprecision	Decreases 1 or 2 levels
Publication bias	Decreases 1 or 2 levels

Table 35.3. Factors That can Increase the Quality of the Evidence

Factor	Consequence
Large magnitude of effect	Increases 1 or 2 levels
All plausible confounding would reduce the demonstrated effect or increase the effect if no effect was observed	Increases 1 level
Dose-response gradient	Increases 1 level

Table 35.4. Canadian Task Force on the Periodic Health Examination's Levels of Evidence

Level	Type of evidence
I	At least 1 RCT with proper randomization
II.1	Well designed Cohort or Case-Control study
II.2	Time series comparisons or dramatic results from uncontrolled trials
III	Expert opinions

Table 35.5. Levels of Evidence from Sackett

Level	Type of evidence
I	Large RCTs with clear cut results
II	Small RCTs with unclear results
III	Cohort and Case-Control studies
IV	Historical Cohort or Case-Control studies
V	Case series, studies with no controls

On the other hand, in **CENETEC's CPGs**, some levels of evidence are graded based in the NICE method, as shown in **Table 35.6**.

Since the introduction of levels of evidence, several other organizations and journals have adopted **variations** of the classification system. Because research questions are divided into the **categories** (treatment, prognosis, diagnosis, and economic/decision analysis), diverse specialties are often asking different types of questions. Therefore, it has been recognized the type and level of evidence needs to be **modified accordingly**, leading to the developed **different types of grading systems**.

From Evidence to Recommendations and their Strength

The **strength of a recommendation** reflects the extent to which a guideline panel is **confident** that desirable effects of an intervention outweigh undesirable effects, or vice versa, across the range of patients for whom the recommendation is intended.

The GRADE approach specifies **two categories** of the strength of a recommendation:

- » **Strong recommendation:** Is one for which the guideline panel is confident that the desirable effects of an intervention outweigh its undesirable effects (strong recommendation for an intervention), or that the undesirable effects of an intervention outweigh its desirable effects (strong recommendation against an intervention).
- » **Weak recommendation:** Is one for which the desirable effects probably outweigh the undesirable effects (weak recommendation for an intervention), or undesirable effects probably outweigh the desirable effects (weak recommendation against an intervention), but appreciable uncertainty exists.

Table 35.6. CENETEC's Levels of Evidence

Level	Type of evidence
Ia	Systematic Review or Meta-analysis of RCTs
Ib	At least one RCT
IIa	At least one well designed and non-randomized controlled study
IIb	At least one well designed quasi-experimental study (cohort study)
III	Well designed and non-experimental studies, descriptive studies, case-control studies, case series
IV	Expert committee reports, opinions and clinical experiences

There are **limitations** to formal grading of recommendations. Like the quality of evidence, the **balance** between desirable and undesirable effects reflects a **continuum**.

In the **CENETEC's CPGs**, some levels of recommendation are graded based in the NICE method, as shown in **Table 35.7**.

In the **UpToDate grading system**, the strength of any recommendation depends on two factors: the **trade off** between benefits and risks and burden, and the **quality of the evidence** regarding treatment effect. The framework summarized in **Table 35.8** generates recommendations from the **very strong** (benefit/risk trade off unequivocal, high quality evidence, 1A) to the **very weak** (benefit/risk questionable, low quality evidence, 2C).

Table 35.7. CENETEC's Levels of Recommendation

Level	Recommendation
A	Based on level I evidence
B	Based on level II evidence or extrapolated from level I evidence
C	Based on level III evidence or extrapolated from level I and level II evidence
D	Based on level IV evidence or extrapolated from level I, level II and level III evidence
A NICE	Recommendation taken from the NICE guide or technology evaluation
GPP	Point of good clinical practice based on clinical experience

Table 35.8. UpToDate's Grading Recommendations

Grade of Recommendation	Clarity of risk/benefit	Quality of supporting evidence	Implications
1A Strong recommendation, high quality evidence	Benefits clearly outweigh risk and burdens, or vice versa.	Consistent evidence from well performed randomized, controlled trials or overwhelming evidence of some other form. Further research is unlikely to change our confidence in the estimate of benefit and risk.	Strong recommendations, can apply to most patients in most circumstances without reservation. Clinicians should follow a strong recommendation unless a clear and compelling rationale for an alternative approach is present.
1B Strong recommendation, moderate quality evidence	Benefits clearly outweigh risk and burdens, or vice versa.	Evidence from randomized, controlled trials with important limitations (inconsistent results, methodological flaws, indirect or imprecise), or very strong evidence of some other research design. Further research (if performed) is likely to have an impact on our confidence in the estimate of benefit and risk and may change the estimate.	Strong recommendation and applies to most patients. Clinicians should follow a strong recommendation unless a clear and compelling rationale for an alternative approach is present.
1C Strong recommendation, low quality evidence	Benefits appear to outweigh risk and burdens, or vice versa.	Evidence from observational studies, unsystematic clinical experience, or from randomized, controlled trials with serious flaws. Any estimate of effect is uncertain.	Strong recommendation, and applies to most patients. Some of the evidence base supporting the recommendation is, however, of low quality.
2A Weak recommendation, high quality evidence	Benefits closely balanced with risks and burdens.	Consistent evidence from well performed randomized, controlled trials or overwhelming evidence of some other form. Further research is unlikely to change our confidence in the estimate of benefit and risk.	Weak recommendation, best action may differ depending on circumstances or patients or societal values.
2B Weak recommendation, moderate quality evidence	Benefits closely balanced with risks and burdens, some uncertainty in the estimates of benefits, risks and burdens.	Evidence from randomized, controlled trials with important limitations (inconsistent results, methodological flaws, indirect or imprecise), or very strong evidence of some other research design. Further research (if performed) is likely to have an impact on our confidence in the estimate of benefit and risk and may change the estimate.	Weak recommendation, alternative approaches likely to be better for some patients under some circumstances.
2C Weak recommendation, low quality evidence	Uncertainty in the estimates of benefits, risks, and burdens; benefits may be closely balanced with risks and burdens.	Evidence from observational studies, unsystematic clinical experience, or from randomized, controlled trials with serious flaws. Any estimate of effect is uncertain.	Very weak recommendation; other alternatives may be equally reasonable.

Key Terms

Define the following terms.

GRADE Approach
Levels of evidence

Quality of evidence
Strength of recommendation

Strong recommendation
Weak recommendation

Active Learning Section

Consolidate the knowledge you acquired in this Chapter through the following exercises.

1. Draw a table summarizing the systems available to determine the level of evidence.
2. Try to answer the following questions:
 - » What is the influence of the evidence in the medical work in Mexico?
 - » Do we understand well what the evidence is, its levels and interpretations?
 - » Which of the evidence classification proposals is most suitable for us?
3. The following clinical settings were obtained from CPGs available at the CENETEC webpage. Interpret the “Levels of Evidence” in parenthesis for each of the clinical settings. Please consider:
E = Evidence
R = Recommendation
 1. E: Transvaginal bleeding in the first trimester of pregnancy, with or without abdominal pain during the early stages of pregnancy, affects between 16 and 25% of all pregnancies (III).
R: In case of transvaginal bleeding of the first trimester, see CPG of abortion (D).
 2. E: In newborns younger than 30 weeks of gestation, exogenous surfactant should be administered prophylactically between 10–30 minutes after resuscitation and neonatal stabilization (1a).
R: Prophylaxis within the first 15 minutes after birth should be administered to almost all preterm infants with respiratory distress syndrome who require intubation for stabilization (A).
 3. E: Detection of prostate specific antigen reduces the mortality rate from prostate cancer by 20%, but is associated with a high risk of overdiagnosis (Ib).
R: It is recommended to practice the study annually from 50 years of age (2a).
 4. E: Starting breast cancer screening at age 40 has been estimated to reduce associated mortality by 14/10,000/year. However, the number of non-cancer biopsies increases (III).
R: Mammography every 2 years is recommended for women between 50–74 years (B).
 5. E: Although the immunity conferred by the measles vaccine is reported to persist for at least 20 years and is believed to last a lifetime in most people, there are no studies to support it (III).
R: It is recommended to apply the first dose of measles vaccine at 12 months of age and the second dose at 6 years of age (A).
 6. E: Stroke is more frequent in men. Men have a higher incidence according to age with the exception of the groups between 35 and 44 years old and over 85 years old (III).

R: Monitoring of Vascular Risk Factors is recommended for people with non-modifiable Risk Factors. And stricter control in elderly patients with a family history of stroke (**GPP**).

7. E: There are insufficient quality data on the risk-benefit of pharmacological treatment of mild depression in the elderly (**IV**).

R: Antidepressants are not recommended in the initial treatment of mild depression in the elderly because their risk-benefit is poor and other therapeutic strategies may be considered (**D**).

8. E: The screening test of choice for cervical cancer is liquid-based cervical cytology (**II**).

R: The optimal age to start screening is unknown and is documented with the natural history of human papillomavirus infection. Screening is recommended to be performed reliably and within three years after the first sexual intercourse or until age 21, whichever occurs first (**2a**).

9. E: The gold standard for the diagnosis of osteoporosis is bone densitometry of the proximal femur and lumbar spine (**I**).

R: All women 65 years of age or older should have bone densitometry regardless of their risk factors (**2**).

10. E: There is no evidence that speaks of how quickly the fluid deficit should be replaced in cases of acute diarrhea in children younger than 5 years (**4**).

R: Rehydration for a period of 4 hours is recommended (**D**).

11. E: Adults over 65 years of age are considered at risk of serious complications from pneumococcal infection (**3**).

R: Vaccination against pneumococcus is recommended in people over 65 years of age with chronic lung diseases, in immunocompromised adults or in long-term care homes, as well as in staff who are in contact with or that take care of them (**B**).

12. E: Eating milk or yogurt one or more times a day is related to lower levels of uric acid compared to those who do not consume them (**2+**).

R: The NSAIDs that have been shown to be more effective in managing gout are Indomethacin, Diclofenac, Naproxen, and Etoricoxib (**C**).

13. E: In pregnant women with Rh negative blood group and not alloimmunized, the application of 300 µg (1500 IU) of anti-D immunoglobulin at 28 weeks significantly reduces the risk of sensitization; 0.2 compared to 1.9% when not administered (**1a**).

R: In RhD negative, non-sensitized women, with a RhD positive newborn and with negative direct coombs test, screening for fetal-maternal hemorrhage should be performed using the rosette test. If negative, administer 300 µg of anti-D immunoglobulin within the first 72 h. of the puerperium (**B**).

14. E: Taking a bath is better to decrease the likelihood of surgical site infection when it is compared to just surgical site cleanliness (**1+**).

R: Recommend to patients that they should take a bath using soap the day before or the day of the surgery (**B**).

15. E: Improving diet represents the most desirable and sustainable method of preventing micronutrient deficiency (**III**).

R: Instruct and encourage parents and staff involved in food preparation to consume those that are rich in iron content (**C**).

Bibliography and Suggested Reading

- Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490.
- Burns PB, Rohrich RJ, Chung KC. The Levels of Evidence and their role in Evidence-Based Medicine. *Plast Reconstr Surg*. 2011 Jul; 128(1): 305–310.
- Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J*. 1979;121:1193–254.
- Peterson PN, Rumsfeld JS. The evolving story of guidelines and health care: does being NICE help? *Ann Intern Med*. 2011;155(4):269.
- Ransohoff DF, Pignone M, Sox HC. How to decide whether a clinical practice guideline is trustworthy. *JAMA*. 2013 Jan;309(2):139–40.
- Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;95:2S–4S.
- Schüenemann H, Brozek J, Guyatt G, Oxman A. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated October 2013. Available at: <https://training.cochrane.org/resource/grade-handbook>.
- Shekelle P. Overview of clinical practice guidelines. Aronson MD, ed. UpToDate. Waltham, MA: UpToDate Inc. <https://www.uptodate.com> (Accessed on May 8th, 2020).
- UpToDate Grading Guide. UpToDate. Waltham, MA: UpToDate Inc. <https://www.uptodate.com> (Accessed on May 8th, 2020).

Appendices

Appendix A	Answers to Active Learning Multiple Choice Questions
Appendix B	Bradford Hill's Criteria for Causality

Answers to Active Learning Multiple Choice Questions

Chapter 8

1. A
2. D
3. B
4. B
5. A
6. B
7. A
8. A
9. B
10. B
11. A
12. C
13. C
14. A
15. B
16. B
17. D
18. C
19. D
20. C

Chapter 9

1. B
2. C
3. C
4. B
5. C
6. B
7. C
8. D
9. C
10. E
11. C
12. A
13. A
14. C
15. D
16. C
17. E
18. B
19. B
20. B

Chapter 10

1. B
2. C
3. B
4. C
5. B
6. B
7. C
8. C
9. B
10. B
11. C
12. D
13. C
14. A
15. D
16. A
17. C
18. B and D
19. A
20. A

Chapter 12

1. C
2. E
3. D
4. E
5. A
6. C
7. B
8. C
9. C
10. C
11. B
12. B
13. C
14. B
15. B
16. E
17. C
18. B
19. D
20. C

Chapter 14

- 1. D
- 2. D
- 3. B
- 4. C
- 5. B

Chapter 15

- 1. B
- 2. B
- 3. C
- 4. C
- 5. A

Chapter 16

- 1. C
- 2. D
- 3. A
- 4. B
- 5. C

Chapter 17

- 1. D
- 2. B
- 3. D
- 4. A
- 5. E
- 6. D
- 7. B
- 8. B
- 9. D
- 10. B
- 11. B
- 12. D
- 13. B
- 14. C
- 15. C

Chapter 19

- 1. C
- 2. C
- 3. A and B
- 4. A
- 5. C
- 6. A
- 7. C
- 8. D
- 9. B
- 10. A

Chapter 20

- 1. D
- 2. D
- 3. C
- 4. C
- 5. D

Chapter 24

- 1. C
- 2. E
- 3. C
- 4. C
- 5. B
- 6. B
- 7. D
- 8. C
- 9. A
- 10. C

Chapter 28

- 1. C
- 2. C
- 3. A
- 4. B
- 5. C
- 6. D
- 7. D
- 8. D
- 9. C
- 10. B
- 11. A
- 12. C
- 13. A and B
- 14. C
- 15. D
- 16. B
- 17. A
- 18. B
- 19. D
- 20. E

Chapter 33

- 1. C
- 2. B
- 3. A
- 4. B
- 5. D

Bradford Hill's Criteria for Causation

Learning objectives for this appendix

- A. Describe the nine Bradford Hill criteria for causality.
- B. Give examples of each of the Hill's criteria.

In 1965, English epidemiologist and statistician, Sir Austin Bradford Hill identified the **nine factors** that constitute the current standards for determining **causation**. Hill's conclusions were developed to answer the question of whether cigarettes cause disease, especially lung cancer.

It is important to note that satisfying these criteria may **lend support** for causality. But failing to meet some criteria, **does not necessarily provide evidence against causality**.

Hill's criteria outline the minimal conditions needed to establish a causal relationship, and they should be viewed as a **guideline**, not as a check list that must be satisfied for a causal relationship to exist. These nine criteria are summarized in **Table AB.1**, and briefly described below.

Strength of Association

Refers to the strength of association between the exposure of interest and the outcome. Is most commonly measured via **risk ratios**, **rate ratios** or **odds ratios**.

Strong associations occur when an exposure is a **strong risk factor**, and there are few other risk factors for the disease.

You **should not assume** that a strong association alone is indicative of causality, as the presence of strong confounding may erroneously lead to a strong causal association.

» **Example:** Bradford Hill pointed out that smoking is a strong risk factor for lung cancer. Smokers are 15 to 30 times more likely to have lung cancer or die due to lung cancer when compared with people who do not smoke. In addition, studies have shown that the risk of lung cancer may be increased 20-fold or more when heavy smokers are compared with non-smokers.

Hill believed that causal relationships were more likely to demonstrate strong associations than were non-causal agents.

Table AB.1. Bradford Hill's criteria of causality

Strength of association	Whether those with the exposure are at a higher risk of developing disease and if so, how much more risk? This criterion suggests that a larger association increases the likelihood of causality.
Consistency	The credibility of findings increases with repetition of findings, including consistency of study findings across different populations and geographical locations.
Specificity of association	Causality is more likely if the exposure causes only one specific disease or syndrome, or if a specific location or population are being affected.
Temporality	This criterion requires that the exposure must occur before the disease, and not after a latency period that is too long. This criterion must always be fulfilled for causality to be concluded.
Biological gradient	The argument for causality is stronger in the presence of a dose–response relationship, where higher or longer exposure leads to an increased risk of disease.
Plausibility	A conceivable mechanism for causation between disease and exposure should exist for there to be a causal relationship.
Coherence	The current association should not contradict any previous knowledge available about the disease and/or exposure.
Experimental evidence	This criterion can involve scientific experiments and addresses the association of exposure with disease. However, 'experiment' relates to the decrease in disease risk when the exposure is removed and often involves animal models.
Analogy	This criterion uses previous evidence of an association between a similar exposure and disease outcome to strengthen the current argument for causation.

» There are also **examples of weak but causal associations**: Exposure to environmental tobacco smoke and lung cancer (RR of 1.2).

Consistency

Refers to the **reproducibility** of study results in various populations and situations. Consistency is generally utilized to **rule out other explanations** for the development of a given outcome.

In general, the greater the consistency, the **more likely a causal association**.

However, the lack of consistency **does not rule out** a causal association, because some effects are only produced under specific combinations of causal components. These conditions may not have been met in some studies of other populations.

» **Example:** Only 10% of heavy smokers develop lung cancer. The other causal components are still being investigated.

Specificity of Association

States that if a **single risk factor** consistently relates to a **single effect**, then it **likely** plays a causal role.

It is important to note that there are **few diseases** that have only one causal agent, and since most diseases are caused by a **constellation of factors**, including poverty, crowding, low immunity, inadequate therapy, and the biological etiology.

» **Example:** an one-to-one relationship exists with certain bacteria and the disease they cause (Tuberculosis).

The specificity of association criterion has also been proven to be **invalid** in a number of instances.

» **Example:** evidence clearly demonstrates that smoking does not lead solely to lung carcinogenesis, but to a myriad of other clinical disorders ranging from emphysema to heart disease.

Some authors feel that specificity of association is the weakest of all criteria and may even be misleading.

Temporality

Has been identified as being the most likely to be the **essential element or condition for causality**.

For an exposure to be causal, its presence must **precede** the development of the outcome. Lack of temporality **rules out** causality.

A temporal relationship is easier to establish in a **cohort study** than in a case-control study or retrospective cohort study.

» **Example:** Administration of insulin precedes a fall in blood glucose levels with a time gap that is consistent with insulin's mechanism of action.

Temporality is the only necessary criterion for causality.

Biological Gradient

Relies on **dose response**, that is: “**the dose of the exposure increases, the risk of disease increases**”.

The presence of the dose-response relationship between an exposure and outcome provides good evidence for a causal relationship. However, its absence **should not be taken as evidence against such a relationship**.

» **Example:** Lung cancer by current amount smoked. Some diseases do not display a dose response relationship with a causal exposure. They may demonstrate a threshold association where a given level of exposure is required for disease initiation, and any additional exposure does not affect the outcome.

The dose response relationship is one of the strongest guidelines, because a confounder is unlikely to cause the same disease gradient as a primary exposure.

Some exposures do not cause disease until the exposure threshold is reached.

» **Example:** skin burns and UV radiation, and cataracts and ionizing radiation.

Plausibility

Generally comes from **basic laboratory science**.

It is not unusual for epidemiological conclusions to be reached in the absence of evidence from a laboratory, particularly in situations where the epidemiological results are the **first evidence** of a relationship between an exposure and an outcome. However, one can further support a causal relationship with the **addition** of a reasonable biological mode of action, even though hard data may not yet be available.

Laboratory experimental evidence increases our confidence in drawing causal conclusions, but **is not essential**.

Arguments about biologic plausibility about an observed exposure response association are too often based only on prior beliefs and the experience of the laboratory scientists.

» **Example:** Environmental tobacco smoke cannot cause lung cancer because the doses are much below those causing cancer in animals.

Coherence

Represents the idea that for a causal association to be supported, any new data **should not be an opposition to the current evidence**. That is, providing **evidence against causality**.

This criterion is more demanding than biologic plausibility in that its evidence must be **extensive** and cutting across disciplinary lines, mutually supporting a causal association between exposure and health outcome.

You should be cautious in making definite conclusions regarding causation, since it is possible that conflicting information is **incorrect or highly biased**.

Another interpretation for coherence can be: when exposure is shown to result in a cluster of related health events.

» **Example:** Smoking causes inflammation of the respiratory tract, release of free radicals, conversion of cells to pre-neoplastic states, transformation of cultured cells to cancer, activation of oncogenes, and lung cancer in humans.

Experimental Evidence

Today's understanding of this criterion results from many areas: the **laboratory, epidemiological studies, preventive, and clinical trials**.

Ideally, experimental evidence must be obtained from a well-controlled study, specifically **randomized clinical trials (RCTs)**. These types of studies can support causality by demonstrating that **altering the cause alters the effect**.

» **Example:** Imagine a clinical trial where researchers control sun exposure to examine effects on skin cancer, randomizing individuals to high sun exposure and some to low sun exposure.

Randomized trials are the **most persuasive studies to establish causality** because they tend to balance unmeasured confounders between exposed and unexposed. However, their use is **limited to risk factors that can ethically be randomized** among subjects.

Analogy

When a factor is suspected of causing an effect, then others factors **similar** or **analogous** to the supposed cause should also be considered and identified as a possible cause, or otherwise eliminated from the investigation.

Analogy is **speculative** in nature, and is dependent upon the **subjective opinion** of the researcher. Therefore is one of the **weakest criteria**.

Absence of analogies **should not** be taken as evidence against causation.

Example: A range of hormones exist which enhance insulin action or produce a similar glucose-lowering effect.

Bibliography and Suggested Reading

- Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol*. 2015; 12: 14.
- Goodman KJ, Phillips CV. *Hill's Criteria of Causation*. Wiley Statistics Reference. 2014;1–4.
- Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295-300.
- Lucas RM, McMichael AJ. Association or causation: evaluating links between "environment and disease". *Bulletin of the World Health Organization*. 2005;83(10):792-95.
- Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health*. 2011;65:392e395.

Glossary

95% confidence interval: Range in which we can be approximately 95% certain that the true population value lies.

Absolute risk reduction (ARR): The reduction in risk (probability of the outcome) that is conferred by the new treatment.

Alpha level: see type I error.

Alternative hypothesis: In statistical hypothesis testing, this is a hypothesis other than the one that is being tested. The alternative hypothesis contains feasible conditions, whereas the null hypothesis specifies conditions that are under test.

Analysis of variance: A method of decomposing the total variability in a set of observations, as measured by the sum of the squares of these observations from their average, into component sums of squares that are associated with specific defined sources of variation.

Analytic epidemiology: Study that involves answering the questions: Why and How. These questions are addressed using hypotheses about relationships and statistical tests for assessing the hypotheses. A comparison group is involved.

Analytic studies: A type of epidemiologic study that tests one or more predetermined hypotheses about associations between exposure and outcome variables. These studies make use of a comparison group.

Arithmetic mean: The measure of central location one is likely most familiar with because it has many desirable statistical properties; it is

the arithmetic average of a distribution of data.

Attack rate: Calculated by dividing the number of cases by the number of people followed. It involves a specific population during a limited time period, such as during a disease outbreak. It is also referred to as a cumulative incidence rate or risk.

Attributable risk: The amount of absolute risk of a health-related state or event among the exposed group that can be attributed to the exposure. It is assumed that the exposure is a cause of the outcome.

Attributable risk percent: Among cases that are exposed, it is the percentage of those cases attributed to the exposure. It is assumed that the exposure is a cause of the outcome.

Available case analysis: Only participants with final study outcomes are included in the data analysis but participants are maintained in the group to which they were allocated. The results may be influenced by bias and confounders.

Average: See Arithmetic Mean.

Background question: Clinical questions about physiology, pathology, epidemiology, and general management, often asked by clinicians in training. Answers to background questions are often best found in textbooks or narrative review articles.

Bar charts: Commonly used for graphically displaying a frequency distribution that involves nominal or ordinal data.

Bayes' theorem: An equation for a conditional probability such as $P(A|B)$ in terms of the reverse conditional probability $P(B|A)$.

Bias: the deviation of the results from the truth; can explain an observed association between exposure and outcome variables that is not real. See also systematic error.

Bimodal distribution: A distribution with two modes.

Biological plausibility: A causal association is consistent with existing medical knowledge.

Blind: Patients, clinicians, data collectors, outcome adjudicators, or data analysts are unaware of which patients have been assigned to the experimental or control group. Also called masking.

Box plot: Also called box and whisker plot. Is a graphical display of data in which the box contains the middle 50% of the data (the interquartile range) with the median dividing it, and the whiskers extend to the smallest and largest values (or some defined lower and upper limits).

Case: A person who has been diagnosed with a health-related state or event.

Case definition: A standard set of criteria applied in a specific situation to ensure that cases are consistently diagnosed, regardless of where or when they were identified and who diagnosed the case.

Case report: A detailed report of the signs, symptoms, diagnosis, treatment, and outcome for an individual patient.

Case series: A group or series of patients that have a similar treatment or exposure. Case reports or case series may provide clues in identifying an adverse health event associated with an exposure.

Case severity: The severity of the illness.

Case-control study: Grouping people as cases (people experiencing a health-related state or event) and controls and investigating whether the cases are more or less likely than the controls to have had past experiences,

lifestyle behaviors, or exposures.

Case-crossover study: Compares the exposure status of a case immediately before its occurrence with that of the same case at a prior time.

Case-fatality rate: The proportion of people with a given disease who die from the disease within a specified time period. This measure is an indicator of the seriousness of the disease and the prognosis for those with the disease.

Categorical data: Data consisting of counts or observations that can be classified into categories. The categories may be descriptive.

Causal inference: A conclusion about the presence of a health-related state or event and the reasons for its existence.

Cause: something that produces an effect, result, or consequence in another factor.

Central limit theorem: The simplest form of the central limit theorem states that the sum of n independently distributed random variables will tend to be normally distributed as n becomes large. It is a necessary and sufficient condition that none of the variances of the individual random variables are large in comparison to their sum. There are more general forms of the central theorem that allow infinite variances and correlated random variables, and there is a multivariate version of the theorem.

Central tendency: The tendency of data to cluster around some value. Central tendency is usually expressed by a measure of location such as the mean, median, or mode.

Censored observations: Used to describe participants who withdraw from the study or who do not experience the outcome of interest.

Chance: A factor to consider when establishing the validity of a statistical association. Chance may explain a relationship between an exposure and disease outcome when the measured association is based on a sample of the population of interest. If everyone in the population is considered, then

chance does not play a role. An association may appear to exist merely because of the luck of the draw—chance. As the sample size increases, the sample becomes more like the population and the role of chance decreases. The degree to which chance variability occurs may be monitored by the p-value.

Chi-square test: A statistic used to test whether the rate of an outcome is significantly different between two or more exposure groups. The test provides a probability that the outcome and the exposure are independent.

Clinical Epidemiology: Focuses specifically on patients and the application of epidemiologic methods to assess the efficacy of screening, diagnosis, and treatment in clinical settings.

Clinical trial: The assignment of an intervention on the individual level and examination of its effects in a clinical setting.

Coefficient of variation: A measure of relative spread in the data; the standard deviation for a set of values is divided by the mean of those values. This measure allows for comparing the variability among two or more sets of data representing different scales.

Coherence: A criterion in causal inference wherein there is consistency with known epidemiologic patterns of disease.

Cohort: A group or body of people, often defined by experiencing a common event (e.g., birth, training, enrollment) in a given time span.

Cohort effect: The change and variation in the health-related state or event of a study population as the study group moves through time.

Confidence interval: A range of reasonable values in which a population parameter lies, based on a random sample from the population.

Confounder: Lurking variable; an extrinsic factor that is associated with a disease outcome and, independent of that association, is also associated with the exposure. Failure to control for a confounder can cause the measured association between exposure and outcome

variables to be misleading.

Confounding: To cause to become confused or perplexed; when the internal validity of a study is compromised because the research failed to control or eliminate a confounder.

Consistency of association: The relationship between an exposure and outcome variable is replicated by different investigators in different settings with different methods.

Contingency table: A tabular arrangement expressing the assignment of members of a data set according to two or more categories or classification criteria.

Continuous data: Information that can take on any value on a continuum or scale.

Control event rate (CER): The frequency of the outcome in the control (current best practice treatment or placebo) group.

Correlation: In the most general usage, a measure of the interdependence among data. The concept may include more than two variables. The term is most commonly used in a narrow sense to express the relationship between quantitative variables or ranks.

Correlation coefficient: A dimensionless measure of the interdependence between two variables, usually lying in the interval from -1 to $+1$, with zero indicating the absence of correlation (but not necessarily the independence of the two variables).

Cross-over trial: A study in which participants receive two or more treatments given consecutively, usually in a random order. The response to the first treatment can be contrasted with the response to the second treatment in the same participants.

Cross-sectional survey: A method of data collection to determine the prevalence of a selected attribute or attributes in a population of interest at a given point in time.

Cumulative incidence rate: A measure of the risk of a health-related state or event in a defined population during a specified time period. Typically calculated by dividing the number of new events in a population by those

at risk of the event at the beginning of the specified time period and multiplied by a rate base of 100. See attack rate.

Degrees of freedom: The number of independent comparisons that can be made among the elements of a sample. The term is analogous to the number of degrees of freedom for an object in a dynamic system, which is the number of independent coordinates required to determine the motion of the object.

Dependent variable: The response variable in regression or a designed experiment.

Descriptive epidemiology: Provides a description of the who (person), what (clinical characteristics), when (time), and where (place) aspects of health-related states or events in a population.

Descriptive study designs: The most common types of descriptive study designs are case reports and case series, cross-sectional surveys, and ecologic designs.

Diagnostic test: Test used to confirm disease in people who present with signs or symptoms.

Discrete data: Integers or counts that differ by fixed amounts, with no intermediate values (e.g., number of people exposed, number of disease cases, number of children).

Dispersion: The amount of variability exhibited by data.

Double-blind study: Neither the participants nor the assessing investigator(s) know who is receiving the active treatment.

Ecologic study: An epidemiologic study where specific individuals are not studied, but instead groups of people are compared, such as comparing injury rates from one occupation to another.

Effect modifier: An extrinsic factor that modifies the association between two other variables.

Effect size: The distance between two mean values, described in units of their standard deviations, that describes the relative

magnitude of the difference between two groups.

Effectiveness: The ability of a program to produce benefits among those who are offered the program (in “real life”).

Efficacy: The ability of a program to produce a desired effect among those who participate in the program compared with those who do not, and under ideal conditions.

Efficiency: Depends on whether a drug is worth its cost to individuals or society. The most efficacious treatment, based on the best evidence, may not be the most cost-effective option.

Epidemiology: The study of the distribution and determinants of health-related states or events in human populations and the application of this study to the prevention and control of health problems.

Estimate: The numerical value of a point estimator. Also called point estimate.

Estimator: A procedure for producing an estimate of a parameter of interest. An estimator is usually a function of only sample data values, and when these data values are available, it results in an estimate of the parameter of interest. Also called point estimator.

Etiology: The science and study of the causes of disease and their modes of operation.

Event: Outcome of interest, which is typically death but can be a non-fatal or favourable outcome, e.g. discharge from hospital.

Experimental event rate (EER): The frequency of the outcome in the experimental (new treatment) group.

Experimental study: A study which is conducted to test the effect of a treatment or intervention.

Explanatory variable: A characteristic that is hypothesised to influence the outcome variable. In clinical studies the explanatory variable is often the group to which patients have been randomised. In cross-sectional and

cohort studies, explanatory variables are often exposure variables.

External validity: Refers to how well the outcome of a study can be expected to apply to other settings. In other words, this type of validity refers to how generalizable the findings are.

False negative: A diagnostic test that indicates that someone does not have a disease when, in fact, he or she does.

False positive: A diagnostic test that indicates that someone has a disease when, in fact, they do not.

Frequency distribution: a completesummaryof the frequencies of the values or categories of a measurement made on a group of people.

Gaussian distribution: Another name for the normal distribution, based on the strong connection of Karl F. Gauss to the normal distribution; often used in physics and electrical engineering applications.

Gold standard: Test regarded as the most accurate method available for classifying people as disease-positive or -negative.

Goodness of fit: In general, the agreement of a set of observed values and a set of theoretical values that depend on some hypothesis. The term is often used in fitting a theoretical distribution to a set of observations.

Hazard ratio: The risk of the event in a study group divided by the risk of the event in a reference group.

Histogram: A frequency distribution for discrete or continuous data.

Hypothesis: A suggested explanation for an observed phenomenon or a reasoned proposal predicting a possible causal association among multiple phenomena.

Hypothesis testing: Any procedure used to test a statistical hypothesis.

Incidence: The number of new cases of a condition that develop in a population during a defined time period.

Incidence density rate: Accounts for varying time periods of follow-up. See also person-time rate.

Incidence rate: Number of new cases of a specified health-related state or event reported during a given time period divided by the estimated population at mid-interval.

Independence: A property of a probability model and two (or more) events that allows the probability of the intersection to be calculated as the product of the probabilities.

Independent variable: The predictor or regressor variables in a regression model.

Indirect causal association: Involves one or more intervening factors and is often much more complicated and difficult to understand than a direct causal association.

Independent t-test: Test to measure whether a continuous outcome variable with a normal distribution is significantly different between two groups, e.g. between male and female or between an intervention and a control group.

Intention-to-treat analysis: All participants are analysed in the group to which they were allocated regardless of subsequent events such as non-compliance or withdrawal from the study. This provides a conservative estimate of treatment effect that is not influenced by confounders.

Intercept: The constant term in a regression model.

Internal validity: Is the extent to which a study establishes a trustworthy cause-effect relationship between a treatment and an outcome.

Interquartile range: The middle 50% of the data; the difference between the third quartile (75th percentile) and the first quartile (25th percentile).

Kaplan–Meier statistic: Statistic used to compare the event rate over time between two or more study groups. Also called a log-rank test.

Kurtosis: A measure of the degree to which an unimodal distribution is peaked.

Likelihood ratio: Probability of a positive test in a person with the disease compared to the probability of a positive test in a person without disease.

Line of best fit: Regression line through a set of data points calculated to minimise the sums of the squared residuals.

Longitudinal data: The same sample of respondents is observed in subsequent time periods.

Loss to follow-up: Circumstance in which researchers lose contact with study participants, resulting in unavailable outcome data on those people. This is a potential source of selection bias in cohort studies.

Measures of central tendency: ways of designating the center of the data. The most common measures are the mean, median, and mode.

Median: The number or value that divides a list of numbers in half; it is the middle observation in the data set.

Misclassification: When the exposure or the status of the health-related state or event is inaccurately assigned. In a case-control study, misclassification results if the exposure status is incorrectly assigned.

Mode: The number or value that occurs most often; the number with the highest frequency.

Multicollinearity: A condition occurring in multiple regression where some of the predictor or regressor variables are nearly linearly dependent. This condition can lead to instability in the estimates of the regression model parameters.

Mutually exclusive events: A collection of events whose intersections are empty.

Natural experiment: An unplanned type of experimental study where the levels of exposure to a presumed cause differ among a population in a way that is relatively unaffected by extraneous factors so that the situation

resembles a planned experiment.

Negative likelihood ratio: How much the odds of the disease decreases when a test is negative.

Negative predictive value: Proportion of test-negative people who do not have the disease.

Nested case-control study: A case-control study nested within a cohort study. Also called a case-cohort study.

Nominal data: Unordered categories or classes (e.g., gender, race/ethnicity, marital status, occupation).

Nonparametric methods: Any method of inference (hypothesis testing or confidence interval construction) that does not depend on the form of the underlying distribution of the observations.

Normal values: Range of values in which the majority of people in a population are expected to lie.

Null hypothesis: A hypothesis stating that there is no difference between the study groups.

Number-needed-to-treat (NNT): The number of people who need to receive a new treatment to prevent one adverse event occurring.

Observational analytic study: A study where the investigator does not manipulate exposure status, but that is designed to test a hypothesis.

Observational exploratory study: A study where the investigator does not manipulate exposure status and not enough information is available to formulate hypotheses.

Observational study: A study which is conducted to measure rates of disease in a population or to measure associations between exposures (risk factors) and disease.

Odds: The probability of an event (p) occurring divided by the probability of that event not occurring ($1-p$).

Odds ratio: Ratio of the odds of the outcome occurring in one group divided by the odds of the outcome occurring in another group.

Ordinal data: The order among categories provides additional information (e.g., stage or grade of cancer).

Outcome variable: The outcome measurement in a study, that is, the variable of interest such as the primary illness or disease status indicator.

Outlier: Data points at the extremities of the range or separated from the normal range of the data values. Data points more than three standard deviations from the mean are usually considered to be outliers.

Overfitting: Adding more parameters to a model than is necessary.

P-value: Probability that a difference between study groups would have occurred if the null hypothesis was true.

Paired t-test: A parametric test that measures whether the means of two related continuous measurements are different from one another, typically measurements taken from the same participants on two occasions.

Parameter: An unknown quantity that may vary over a set of values. Parameters occur in probability distributions and in statistical models, such as regression models.

Parametric statistics: Statistics used when the outcome measurement has a distribution that is approximately normal.

Percentile: The set of values that divide the sample into 100 equal parts.

Period prevalence: probability that an individual has been affected by a given disease during a defined time period.

Phase I trial: Initial trial of a new treatment to assess safety and feasibility in a small group of volunteers who do not have the disease or patients with symptoms.

Phase II trial: A clinical trial to measure efficacy, that is, the effect of a treatment under ideal conditions, in patients with the disease.

Phase III trial: Large randomised controlled trial or multi-centre study to measure effectiveness in the community, that is, the effect of a treatment in general clinical practice.

Phase IV surveillance: Post-marketing survey to measure rare adverse events.

Placebo: An inactive substance or treatment given to satisfy a patient's expectation of treatment.

Placebo effect: The effect on patient outcomes (improved or worsened) that may occur because of the expectation by a patient (or provider) that a particular intervention will have an effect.

Point prevalence proportion: All existing cases of the disease or event at a point in time divided by the total study population at the point in time.

Point source: Epidemic in which persons are exposed to the same exposure over a limited time period.

Population: Any finite or infinite collection of individual units or objects.

Population attributable risk: Amount of absolute risk of a health-related state or event in a population that can be attributed to the exposure. This measure assumes that the exposure causes the outcome.

Population attributable risk percent: The percent of the absolute risk of a health-related state or event in a population that can be attributed to the exposure. This measure assumes that the exposure causes the disease.

Positive likelihood ratio: How much the odds of the disease increase when a test is positive.

Positive predictive value: Proportion of test-positive people who have the disease.

Power: The power of a statistical test measures the test's ability to reject the null hypothesis when it is actually false; power is directly associated with sample size. It is equal to $1 - \beta$.

Prediction: The process of determining the value of one or more statistical quantities at some future point in time. In a regression model, predicting the response y for some specified set of regressors or predictor variables also leads to a predicted value, although there may be no temporal element to the problem.

Predictive value negative: The predictive value of a negative is the probability that an individual with a negative test does not have the disease.

Predictive value positive: The predictive value of a positive test is the probability that an individual with a positive test actually has the disease.

Prevalence: The total number of people in a population with a condition at a given point in time.

Prior probability: Prevalence proportion of disease used in calculating the predictive value positive and predictive value negative proportions.

Probability: A numerical measure between 0 and 1 assigned to events in a sample space. Higher numbers indicate the event is more likely to occur.

Prognosis: The prospect of recovery as anticipated from the usual course of disease; a prediction of the probable course and outcome of a disease.

Prognostic indicators: Clinical and laboratory information that help forecast the likely outcome of a disease.

Proportion: A ratio in which the numerator is included in the denominator.

Prospective cohort study: An analytic epidemiologic study that classifies participants according to exposure status and then follows them over time to determine if the rate of developing a given health-related state or event is significantly different between the exposed and the unexposed groups.

Protocol: A detailed written plan of the study; the outline of the study protocol may include the research questions, background and

significance, design (time frame, epidemiologic approach), subjects (selection criteria, sampling), variables (predictor variables, confounding variables, outcome variables), and statistical issues (hypotheses, sample size, and analytic approach).

Qualitative data: Data derived from nonnumeric attributes, such as sex, ethnic origin or nationality, or other classification variable.

Quantiles: The set of $n - 1$ values of a variable that partition it into a number n of equal proportions. For example, $n - 1 = 3$ values partition data into four quantiles with the central value usually called the median and the lower and upper values usually called the lower and upper quartiles, respectively.

Quantitative data: Data in the form of numerical measurements or counts.

Quartile(s): The three values of a variable that partition it into four equal parts. The central value is usually called the median and the lower and upper values are usually called the lower and upper quartiles, respectively. See also Quantiles.

r-value: Pearson's correlation coefficient that measures the strength of a linear relationship between two continuous normally distributed variables.

r squared: The coefficient of determination is equal to the squared correlation coefficient and provides an estimate of the per cent of variation in one variable that is explained by the other variable.

Random: Nondeterministic, occurring purely by chance, or independent of the occurrence of other events.

Random assignment: The random allocation of participants to one or another of the study groups. Participants have an equal probability of being assigned to any of the groups. This process minimizes any confounding effects by balancing out the potential confounding factors among the groups.

Random error: Chance variability; the greater the error, the less precise the measurement.

Random selection: Sample taken from a population in which all people have an equal chance of being selected.

Randomization: A set of objects is said to be randomized when they are arranged in random order.

Randomised controlled trial: A study which is conducted to measure whether a new treatment is superior or equivalent to no treatment or an existing treatment and in which participants are randomly allocated to the study groups.

Range: The difference between the largest (maximum) and smallest (minimum) values of a frequency distribution.

Rank: In the context of data, the rank of a single observation is its ordinal number when all data values are ordered according to some criterion, such as their magnitude.

Rate: A proportion with the added dimension of time. The numerator consists of health-related states or events during a given time period and the denominator consists of persons at risk during the same time period.

Rate ratio: A measure of the strength of association between dichotomous exposure and outcome variables that involves the ratio of person-time rates.

Ratio: A relationship between two quantities, normally expressed as the quotient of one divided by the other.

Recall bias: A type of observation bias (or measurement bias) that can occur in case-control and cross-sectional studies because of differential recall about past exposure status between those who have the disease compared with those who do not. In general, cases tend to have better recall.

Regression: The statistical methods used to investigate the relationship between a dependent or response variable y and one or more independent variables x . The independent variables are usually called regressor variables

or predictor variables.

Regression coefficient(s): The parameter(s) in a regression model.

Regression line: A graphical display of a regression model, usually with the response y on the ordinate and the regressor x on the abscissa. Also called regression curve.

Relative frequency: Derived by dividing the number of people in a group by the total number of people; that is, a part of the group is expressed relative to the whole group.

Relative risk: Ratio of the probability of the outcome occurring in the exposed group divided by the probability of the outcome occurring in the non-exposed group.

Residuals: Distance between an observed value and its predicted value, in this case the value predicted by the regression line.

Retrospective cohort study: An analytic epidemiologic study where the cohort represents a historical cohort assembled using available data sources.

Risk: The probability of an event or outcome occurring, such as the risk of an infection, death or cure.

Risk factor: A factor that is associated with an increased probability of experiencing a given health problem.

Risk ratio: A measure of the strength of association between dichotomous exposure and outcome variables that involves the ratio of attack rates (also called relative risk).

Sample: A subset of items that have been selected from the population.

Screening test: Test used for early identification of disease in a population without symptoms.

Selection bias: Systematic error that occurs from the way the participants are selected or retained in a study (e.g., Berkson's bias in case-control studies and loss to follow-up in cohort studies).

Sensitivity: Proportion of disease-positive people who are test-positive.

Significance: In hypothesis testing, an effect is said to be significant if the value of the test statistic lies in the critical region.

Single-blinded study: A placebo-controlled study in which the subjects are blinded, but investigators are aware of who is receiving the active treatment.

Skewness: A term for asymmetry usually employed with respect to a histogram of data or a probability distribution.

Specificity: Proportion of disease-negative people who are test-negative.

Standard deviation (SD): A measure of variability that describes how far the data spreads on either side of the central mean value. The standard deviation is the square root of the variance and therefore is in the same units as the data values.

Standard error (SE): A measure of the precision with which the mean value has been measured.

Standardize: The transformation of a normal random variable that subtracts its mean and divides by its standard deviation to generate a standard normal random variable.

Statistic: A summary value calculated from a sample of observations. Usually, a statistic is an estimator of some population parameter.

Statistical inference: An inference or conclusion made about a population based on sampled data.

Statistics: The science of collecting, analyzing, interpreting, and drawing conclusions from data.

Strength of association: A critical criterion in causal inference; a valid statistical association and the stronger the strength of that association provides support for the possibility of there being a causal association.

Study design: The plan that directs the researcher along the path of systematically collecting, analyzing, and interpreting data.

Surveillance: Close observation and monitoring of environmental exposures, individuals and communities at risk, outcomes, and so forth.

Survival rate: Proportion of persons in a study or treatment group surviving for a given time after diagnosis.

Survival time: The percent of people who survive a disease for a specific amount of time.

Systematic error: Bias that occurs from differences between the truth addressed by the research question and the subjects and measurements in the study. Recall bias in a case-control study is an example of systematic error, where the cases or controls tend to misclassify their exposure status at different levels.

t-test: Any test of significance based on the t distribution. The most common t-tests are (1) testing hypotheses about the mean of a normal distribution with unknown variance, (2) testing hypotheses about the means of two normal distributions and (3) testing hypotheses about individual regression coefficients.

t-value: A t value, which is calculated by dividing a mean value by its standard error, gives a number from which the probability of the event occurring is estimated from a t-distribution. A t-distribution is closely related to a normal distribution but depends on the number of cases in the sample.

Temporality: A linear process of past, present, and future.

Test statistic: A function of a sample of observations that provides the basis for testing a statistical hypothesis.

Therapeutic trial: A trial used to test new treatment methods. See also clinical trial.

Treatment received analysis: Participants are re-grouped according to the treatment they actually received irrespective of the treatment to which they were allocated. Using this method, there is no control of confounders.

True negative: A negative test result for someone without the disease.

True positive: A positive test result for someone with the disease.

Type I error: A difference between groups is statistically significant although a clinically important difference does not exist. In this case, the null hypothesis is incorrectly rejected. That is, a difference between groups is statistically significant although a clinically important difference does not exist.

Type II error: A difference between groups is not statistically significant although a clinically important difference exists. In this case, the null hypothesis is incorrectly accepted.

Unpaired z-test: Test used to compare the mean values of two independent samples using a normal distribution. This test is only used when the sample size is very large or the mean and standard deviation of the population are known.

Validity: See internal validity and external validity.

Variable: A characteristic that varies from one observation to the next and can be measured or categorized.

Variance: A squared term that describes the total variation in the sample.

Wilcoxon signed rank test: A distribution-free test of the equality of the location parameters of two otherwise identical distributions. It is an alternative to the two-sample t-test for nonnormal populations.

Index

Symbols

2 × 2 contingency table 159,
161, 184, 201, 206
 α 80, 81, 82, 83, 84, 87
 β 83, 84, 87

A

Absolute effect measure
285
Absolute risk reduction 179
Accuracy 205, 262, 268
Active control 252
Active Learning 4, 6, 7, 8
Adaptive design 241
Adaptive Treatment Designs
235
AGREE II 295, 296, 297
Alpha 84
Alternate hypothesis 84
Ambispective cohort study
125, 126
Ambispective study 120
Analysis of variance 90, 94,
96, 97, 100
Analytical study 120, 220
Another active treatment
241
ANOVA 91, 94, 95, 96, 97
Answerable questions 15
Application 5
Appraisal of Guidelines for
Research & Evaluation
295

Area under the ROC curve
215
Assay sensitivity 252
Assimilation 5
Assumption 84
Attention 5
Attributable risk 151, 154,
175, 176, 177, 179
Attributable risk percent 177,
179
Attrition bias 268
AUROC 211, 212, 215
Australian National Health
and Medical Research
council 290

B

Background clinical ques-
tion 21
Bar chart 65, 70
Bayesian approach 196
Baye's theorem 198
Belmont Report 222, 223,
226
Beneficence 226
Benefit 255, 260
Beta 84
Bias 135, 228, 230, 232,
237, 238, 239, 241,
243, 244, 245, 261,
262, 263, 264, 265,
266, 267, 268, 269,
280, 282, 283, 284,

285

Binary numbers 47, 50
Bioequivalence studies 248
Biostatistics 39, 40
Blinding 238, 239, 241, 243,
245
Boolean operators 23, 26,
27, 28
Box-and-whiskers plot 55,
66, 70

C

Case-control Nested Within
a Cohort 131, 132
Case-control study 160,
163, 236
Cases 130, 132
Causation 105, 109
CENETEC 288, 289, 296
Censored data 147
Center 58, 70
Centiles 61, 70
Central limit theorem 70
Central tendency 70
Centro Nacional de Excel-
encia Tecnológica en
Salud 288, 297
Chance 261, 268
Chi-squared test 100
Choice of non-inferiority
margin 252
Clinical architecture 11
Clinical epidemiology 11

Clinical evidence 16, 21
 Clinical practice guidelines 289, 296
 Clinical question 17, 18, 19, 21
 Cohort 123, 124, 125, 126, 127
 Cohort and case studies 132
 Cohort study 123, 124, 125, 126, 127
 Comparison 18, 19
 Comparison group 241
 Comprehension and internalization 5
 CONBIOÉTICA 224, 226
 Concepts 24, 30
 Confidence 77, 84
 Confidence interval 70, 161
 Confidence limits 70, 254
 Conflicts of interest 296
 Confounder 168, 171
 Confounder control 171
 Confounders 261
 Confounding 167, 168, 171, 173, 228, 237, 244
 Constancy and metrics 252
 Continuous measurement 49
 Control group 241
 Controls 130, 131, 132
 Correlation 103, 104, 105, 106, 109, 113
 Cost-effectiveness 260
 Cox Proportional Hazards Model 145
 Cox regression 145, 147
 Crossover design 241
 Cross-sectional studies 135, 136, 139
 Cross-sectional study 120
 Cross-tabulation 70

D

Data 43, 47, 50
 Database 23, 25, 26, 27
 Decision-making 11
 Declaration of Helsinki 222, 224, 226
 Dependent variable 45, 51, 52, 53
 Descriptive statistics 44, 51, 70
 Descriptive study 120
 Descriptive table 70
 Detection bias 265, 268
 Diagnostic accuracy 191
 Diagnostic reasoning 195, 198
 Diagnostic threshold 191
 Dichotomous measure 201
 Difference in incidences 125
 Discrete measurement 49
 Discrimination capacity 215
 Dispersion 58, 60, 70
 Double-blinded 239
 Double dummy 241

E

Early termination of a RCT 241
 Effectiveness 255, 257, 260
 Effect measure 285
 Effect measures 151, 153
 Efficacy 241, 255, 256, 260
 Efficiency 255, 258, 260
 End-point selection 252
 Enriched enrollment design 241
 Epidemiological measures 154
 Equipoise 228, 241
 Equivalence 247, 248, 249, 254
 Equivalence study 241
 Equivalence trials 247, 248,

254

Error 262, 268
 Estimate 42
 Estimation 77, 84
 Ethical requirements 220
 Evaluation 5
 Evidence-based medicine 11
 Evidence pyramid 285
 Evidence synthesis 273, 276
 Exclusion criteria 135
 Experimental study 120, 220
 External validity 237, 241, 256, 260, 263, 280, 285

F

Factorial design 241
 Fagan nomogram 185, 191
 False negative 201, 202, 206
 False-negative 84
 False positive 201, 202, 206
 False-positive 84
 Fisher's exact test 90, 100
 Foreground clinical question 21
 Fraction attributable 125
 Frequency bar chart 70
 Frequency distribution 70
 Friedman test 90, 97, 100

G

Generalizability 260
 Gold standard 206
 GRADE Approach 299, 305
 Group-randomized trial design 241

H

Hazard 145, 146, 147, 150
 Hazard rate 147, 150

Hazard ratio 147, 150
 Hierarchy of evidence 11
 Histogram 64, 70
 Homogeneity of variance
 78, 84
 Hypothesis 77, 78, 79, 82,
 84, 85
 Hypothesis testing 84, 85

I

Imprecision 268
 Incidence 125, 126
 Incidence-prevalence bias
 266, 268
 Incidence rate 152, 154
 Inclusion criteria 135
 Independence of the data
 78, 84
 Independent censoring 147
 Independent variable 44,
 45, 51, 52
 Inference 44
 Inferential statistics 43, 44,
 52
 Information bias 266, 268,
 269
 Informed consent 222, 233,
 241
 Inaccuracy 268
 Integers 48
 Intention-to-treat 260,
 265
 Interest 5
 Internal validity 237, 241,
 256, 260, 263, 280,
 285
 Inter-quartile range 70
 Interval 48, 50, 52
 Intervention 18, 19, 230, 241

J

Justice 223, 224, 226

K

Kaplan-Meier curve 143,
 144, 145, 147
 Kaplan-Meier estimator 144
 Keywords 24, 26, 30
 Knowledge acquisition 5
 Kruskal-Wallis test 90, 94,
 96, 100
 Kurtosis 62, 70

L

Learning 3, 4, 5, 6, 7, 8
 Legal framework 226
 Length of follow-up 147
 Levels of evidence 300,
 305
 Likelihood 183, 186, 187,
 188, 191, 192
 Likelihood ratio 191
 Limits to the search strategy
 30
 Location of evidence syn-
 thesis 276
 Log rank test 90
 Log-rank test 145
 Longitudinal study 120
 Loss-to-follow-up bias
 266, 268
 LR of a negative test 185
 LR of a positive test 185

M

Manipulated variable 45
 Mann-Whitney U test 90,
 100
 Masking 238, 241
 Matching 170, 171
 McNemar test 90, 100
 Mean 63, 70, 73
 Measures of implication
 151, 153
 Measures of importance
 151, 153

Median 59, 63, 70, 73
 Medical Subject Headings
 18
 Memory 4, 5
 MeSH terms 18
 Meta-analysis 278, 285,
 286
 Method of least squares
 109
 Missclassification bias 268
 Mode 60, 70, 73, 74
 Motivation 5
 Multivariate models 170, 171

N

Narrative review 285
 Narrative reviews 277
 National Institut for Health
 and Care Excellence
 290
 Negative likelihood ratio 191
 Negative predictive value
 206, 207
 NOM-12-SSA3-2012 224,
 226
 Nominal 48, 50, 52
 Non-inferiority margin 254
 Non-inferiority study 241
 Non-inferiority trials 247,
 254
 Nonparametric 70
 Nonparametric tests 91
 Normal distribution 70
 Null hypothesis 79, 84
 Number needed to treat
 179, 180
 Nuremberg Code 221, 226

O

Observational studies 261
 Observational study 120,
 171, 172, 173
 Observer bias 267, 268
 Odds 157, 160, 161, 164,

165
 Odds ratio 132, 136, 151,
 154, 160, 161, 164
 Open label 239
 Ordinal 48, 50, 52
 Outcome 18, 20
 Outcome of Interest 142
 Overall odds of death 161
 Overall risk of death 161

P

Paired data 91, 100
 Parallel groups design 241
 Parametric 70
 Parametric tests 91
 Pearson correlation 103,
 104, 109
 Pearson correlation test
 100
 Performance bias 268
 Period prevalence 153, 154
 Per-protocol 260
 Phase III studies 232, 241
 Phase II studies 231, 241
 Phase I studies 241
 Phase IV studies 232, 241
 PICO 15, 17, 18, 19, 20, 21,
 22
 Pie chart 70
 Placebo 229, 241
 Point prevalence 153, 154
 Population 18, 19, 43, 50
 Population attributable risk
 151, 154, 177, 178, 179
 Population attributable risk
 percent 178, 179
 Population parameter 43,
 44, 51, 52
 Positive likelihood ratio 191
 Positive predictive value
 206, 207
 Post-marketing surveillance
 241
 Post-test odds 195, 198

Post-test probability 183,
 185, 186, 188, 195,
 198
 Power 145, 183
 Precision 262, 268
 Predictive value 206
 Preferred Reporting Items
 for Systematic Re-
 views and Meta-Anal-
 yses 280, 285
 Pre-test odds 195, 198
 Pre-test probability 183,
 185, 186, 187, 188,
 189, 195, 198
 Prevalence 136, 151, 152,
 153, 154
 Primary questions 17, 21
 Primary studies 276, 285
 PRISMA 280, 285, 286
 Probability 84
 Problem 18, 19
 Product limit estimator 147
 Prognosis 16, 20
 Proportion attributable 125,
 126
 Proportion of cases and
 controls 132
 Prospective cohort studies
 159
 Prospective cohort study
 124, 126
 Prospective study 120, 220
 Publication bias 266, 268
 p-value 77, 81, 82, 83, 84,
 86, 87, 88

Q

Qualitative variables 47
 Quality of evidence 300,
 305
 Quantitative variables 47

R

R2 108, 109

Random error 262, 268
 Randomization 169, 171,
 239, 240, 241, 245
 Randomized clinical trial
 220
 Randomized Clinical Trial
 241
 Randomized clinical trials
 255
 Range 61, 70, 73
 Ratio 48, 50
 Real numbers 48, 50
 Recall bias 267, 268
 Receiver operating charac-
 teristic curve 215,
 216
 Regression 103, 106, 107,
 108, 109, 112, 113
 Regression model 109
 Rejecting hypothesis 84
 Related terms 25, 30
 Relative effect measure 285
 Relative risk 125, 126, 151,
 154, 161, 164
 Reporting bias 267, 268
 Reproducibility 10, 11
 Research hypothesis 79, 84
 Residuals 109
 Respect for Persons 223,
 226
 Restriction 170, 171
 Retrospective cohort study
 125, 126
 Retrospective study 120
 Right censoring 147
 Risk 157, 159, 161, 164, 165
 ROC curve 211, 212, 213,
 214, 215

S

Sample 44, 50, 53
 Sample statistic 43, 44
 Sampling error 44, 51
 Scale of measure 70

Scottish Intercollegiate
 Guidelines Network
 290
 Search strategy 23, 24, 28,
 30, 31
 Secondary questions 21
 Selection bias 239, 244,
 264, 265, 268, 269
 Selection criteria 233, 241
 Sensitivity 83, 184, 186,
 211, 212
 Sensivity 206
 Sequential trial design and
 interim analysis 241
 Shape 58, 62, 70
 Significance 70
 Simpson's paradox 170, 171
 Single-blinded 239
 Skewness 62, 70
 Spearman correlation test
 100
 Spearman's correlation 105,
 109, 111
 Specificity 83, 186, 201,
 202, 203, 204, 205,
 206, 207, 209, 211
 Spectrum of probability 198
 Standard deviation 70, 73,
 74
 Standard error 70
 Statistically significant 70
 Statistical power 84
 Statistical significance 84,
 88
 Statistical test 100
 Statistics 41, 42
 Stopping rule 241
 Stratification 170, 171
 Strength of recommenda-
 tion 305
 Strong recommendation
 302, 304, 305
 Student's t-test 90, 94, 95,
 100
 Study design 120

Study validity 232, 237
 Subject headings 25, 30
 Superiority study 241
 Superiority trials 247, 254
 Survival analysis 141, 147,
 150
 Survival function 143, 147
 Survival time 142, 147
 Symmetry 58, 62, 70
 Synonyms 24, 25, 30
 Systematic error 237, 268
 Systematic review 276, 278,
 285, 286

T

Test statistic 84
 Test threshold 191
 The Guidelines International
 Network 290
 Time-to-Event Analysis
 145, 147
 Transfer 5
 Transverse study 120
 Treatment threshold 191
 Triple-blinded 239
 True negative 201, 202,
 206
 True positive 201, 202, 206
 Type I error 81, 83, 84, 86,
 87, 212, 259
 Type II error 77, 83, 84, 86,
 212, 259

U

Unblinded 239
 Uncertainty 70, 84
 Uncertainty principle 241
 Universe 43, 44
 Unpaired data 100
 Uses of evidence synthesis
 276
 Usual care 241
 Utility 255, 260

V

Validity 145, 201, 206, 232,
 237, 263
 Variability 70
 Variable 43, 44, 45, 46, 47,
 51

W

Weak recommendation
 302, 304, 305
 Wilcoxon rank-sum test 90,
 94
 Wilcoxon sign-rank test 90
 Wilcoxon's rank sum 90,
 100

This book was finished printing in the
Graphic Workshops of the
Universidad Autónoma de San Luis Potosí
in **January 2021**
with a circulation of 500 copies