

Introduction to **Research** in the **Health Sciences**

Stephen Polgar | Shane Thomas

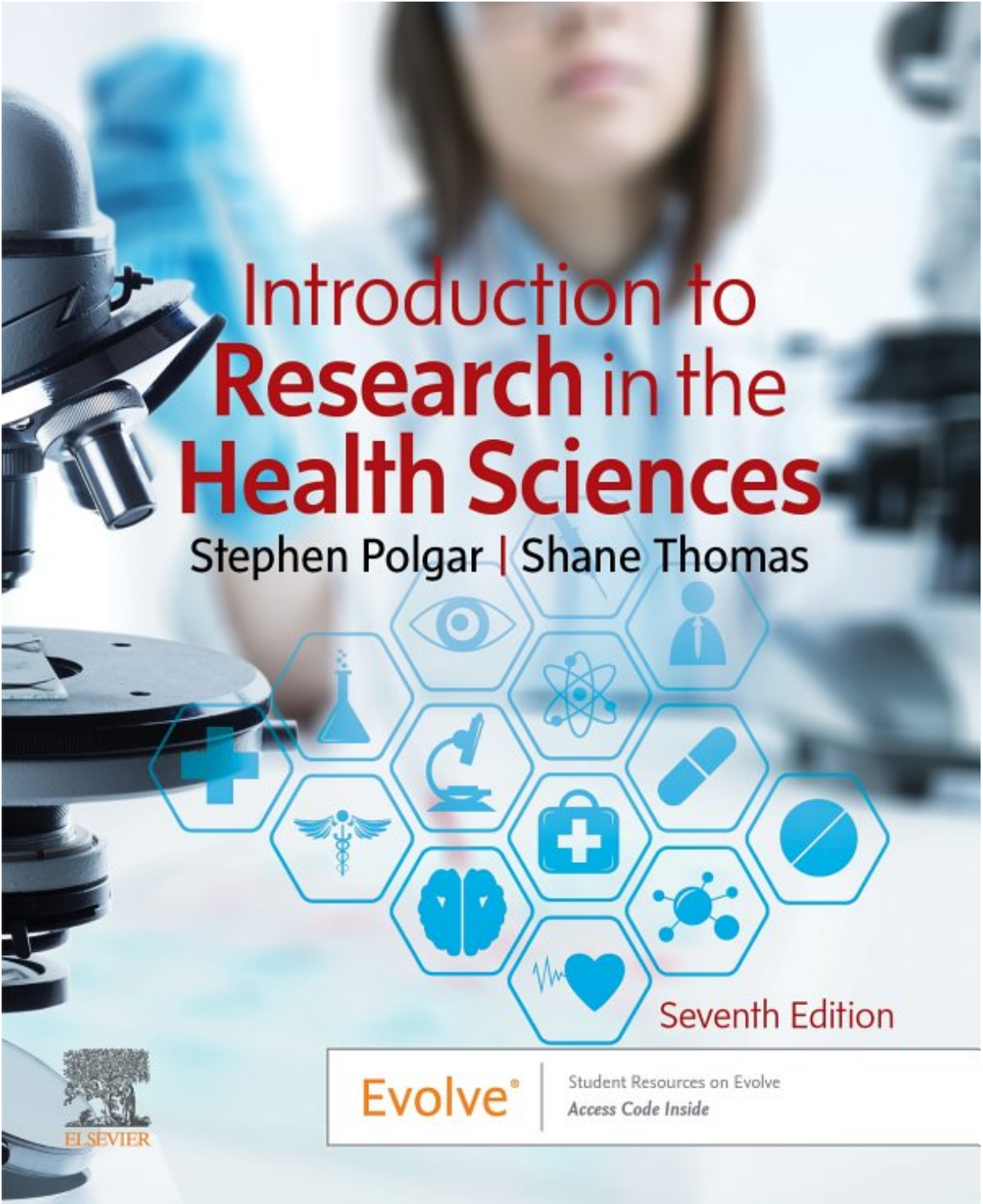


Seventh Edition



Evolve®

Student Resources on Evolve
Access Code Inside



Introduction to **Research** in the **Health Sciences**

Stephen Polgar | Shane Thomas



Seventh Edition



Evolve®

Student Resources on Evolve
Access Code Inside

Introduction to Research in the Health Sciences

SEVENTH EDITION

Stephen Polgar, BSc (Hons) MSc PhD

Retired Senior Lecturer, School of Public Health, Faculty of Health Sciences, La Trobe University, Melbourne, Australia

Shane Thomas, PhD PLD MAPS FCHP

Honorary Professor, Research School in Population Health, Australian National University, Director, International Primary Health Care Research Institute, Shenzhen, People's Republic of China



ELSEVIER Edinburgh London New York Oxford Philadelphia St Louis
Sydney 2020

Table of Contents

Cover image

Title page

Dedication

Copyright

Preface

Section 1. Methodological Foundations of Health Research

Introduction

1. Foundations of Health Research

Introduction

Knowledge and Methods

The Scientific Method

Positivist Paradigm

Postpositivist Paradigms

Pragmatism: The Combined use of Quantitative and Qualitative Methods

Summary

2. Quantitative and Qualitative Methods

Introduction

The Social Construction of Reality

Contrasting Qualitative and Quantitative Methods

Summary

3. The Research Process

Introduction

Sequential Steps of the Research Process

The Structure of a Research Paper

The Publication Process

Research and Evidence-Based Health Care

Summary

Section 2. Research Planning

Introduction

4. The Formulation of Research Questions

Introduction

Identifying Research Problems

Literature Reviews and Research

Formulating the Research Question

Methodological Considerations

Ethical Issues and Procedures

Aims and Hypotheses

Research Proposals

Summary

5. Sampling Methods and External Validity

Introduction

Basic Issues in Sampling

Representative Samples

Sample Size

Sampling in Qualitative Research

Purposive Sampling

External Validity and Sampling

Summary

6. Ethics

Introduction

Philosophical Principles

Resources

Summary

Section 3. Research Designs

Introduction

7. Experimental Designs and Randomized Controlled Trials

Introduction

The Concept of Causality

Confounding and Bias in Research Studies

The Use of Control Groups in Applied Health Research

Assignment of Participants into Groups

Types of Experimental Designs

Randomized Controlled Trials

Controlled Research Involving Human Participants

The External Validity of Experiments And Rcts

Summary

8. Surveys and Quasi-Experimental Designs

Introduction

Naturalistic Non-Rct Research Designs

Surveys

Epidemiology

Designs for Comparing Groups

Quasi-Experimental Research Designs

The Internal and External Validity of Naturalistic Non-Rct Designs

Summary

9. Qualitative Research

Introduction

Research Problems and Questions

Approaches to Qualitative Research

What defines qualitative research?

An Example of Qualitative Research

Summary

Section 4. Data Collection

Introduction

10. Questionnaires and Survey Design

Introduction

Questionnaire Construction

Questions and Questionnaire Formats

The Overall Structure of Questionnaires

Summary

11. Mixed Methods

Introduction

Perspectives to Health Problems: an Example

Quantitative/Qualitative Designs

Summary

12. Interviewing Techniques

Introduction

Interviewing Models

Methods of Conducting Interviews

The Interview Process

Methods of Recording Interview Information

The Analysis of Interview Data

Summary

13. Measurement and Observation

Introduction

Observation

Observation and Measurement

Operational Definitions

How do we operationalize constructs?

Measurement Tools and Procedures

Standardized Measures and Tests

What constitutes a good assessment tool?

Measurement Scale Types

Levels of Measurement

Summary

Section 5. Descriptive Statistics

Introduction

14. Organization and Presentation of Data

Introduction

The Organization and Presentation of Nominal (Categorical) or Ordinal (Ranked) Data

Organization and Presentation of Interval or Ratio Data

Descriptive Statistics for Discontinuous Data

Summary

15. Measures of Central Tendency and Variability

Introduction

Measures of Central Tendency

Comparison of the Mode, Median and Mean

Measures of Variability

Summary

16. Standard Scores and Normal Distributions

Introduction

Standard Scores (z-Scores)

Normal Distributions

Calculations of Areas Under the Normal Curve

Standard Normal Curves for the Comparison of Distributions

Summary

17. Correlation

Introduction

Correlation

Correlation Coefficients

Uses of Correlations in the Health Sciences

Correlation and Causation

Summary

Section 6. Data Analysis and Inference

Introduction

18. Probability and Confidence Intervals

Introduction

Probability

Sampling Distributions

Sampling Distribution of the Mean

Calculating Confidence Intervals

Confidence Intervals Using the t Distribution

Confidence Intervals in Health Research

Summary

19. Hypothesis Testing

Introduction

The Logic of Hypothesis Testing

Steps in Hypothesis Testing

Illustrations of Hypothesis Testing

The Relationship Between Descriptive and Inferential Statistics

Selection of the Appropriate Inferential Test

The χ^2 Test

χ^2 and Contingency Tables

Statistical Packages

Summary

20. Effect Size and the Interpretation of Evidence

Introduction

Effect Size

Confidence Intervals and Effect Size

How to Optimize Statistical Inferences

How to Interpret Null (Nonsignificant) Results

Statistical Power Analysis

Clinical Decision Making

Overview of Conducting Data Analysis

Summary

21. Qualitative Data Analysis

Introduction

Understanding Meaning in Everyday Life

Management and Organization of Data

Coding Qualitative Data

Content Analysis

Thematic Analysis, 'Verstehen' and Grounded Theory

Interpretation and Social Context

The Accuracy of Qualitative Data Analysis

Summary

22. Critical Evaluation of Published Research

Introduction

Guidelines for the Critical Appraisal of Research Publications

Critical Evaluation of the Introduction in a Research Paper

Critical Evaluation of the Methods Section in a Research Paper

Critical Evaluation of the Results Section in a Research Paper

Critical Evaluation of the Discussion Section in a Research Paper

Summary

Section 7. Evaluation and Dissemination of Research Results

Introduction

23. Synthesis of Research Evidence: Systematic Reviews and Meta-Analyses

Introduction

Basic Principles

Narrative Reviews of Research Evidence

Systematic Reviews

Meta-Analysis

Validity of Systematic Reviews

Summary

24. Translational Research

Introduction

(1). Basic Research

(2). Clinical Research

(3). Translational Research

Why Translational Research?

Phases of Clinical Research

‘Lost in Translation’: Example of Obstacles to Successful Translational Research

Summary

A APPENDIX. z-Scores and Associated Areas Between z and Mean and Beyond

Appendix B. t Distribution

APPENDIX C. Chi-square (χ^2)

Index

Dedication

For Sue, Sharon and Rebecca

In appreciation of the support of my loving family and all of the many research colleagues and students with whom I have worked.

Copyright

© 2020, Elsevier Limited. All rights reserved.

First edition 1988

Second edition 1991

Third edition 1995

Fourth Edition 2000

Fifth Edition 2008

Sixth Edition 2013

Seventh Edition 2020

The right of Stephen Polgar and Shane Thomas to be identified as authors of this work has been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds or experiments described herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made. To the fullest extent of the law, no responsibility is assumed by Elsevier, authors, editors or contributors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-7020-7493-6

Content Strategist: Poppy Garraway Smith/Serena Castelnovo

Content Development Specialist: Carole McMurray

Content Coordinator: Susan Jansons/Kirsty Guest

Publishing Services Manager: Shereen Jameel

Project Manager: Radhika Sivalingam

Design: Amy Buxton

Illustration Manager: Anitha Rajarathnam

Marketing Manager: Deborah Watkins

Printed in China

Last digit is the print number: 9 8 7 6 5 4 3 2 1



Preface

Why do we need to know about research to be an effective health professional?

Welcome to the 7th edition of our text! In this book our goal is to give you an effective working knowledge of health research needed to support high quality clinical practice. A key foundation of clinical practice is the use of research evidence concerning what works and with whom in health care settings. This evidence is generated by health research. Hence a basic knowledge of research methods is needed to be an effective clinician.

In this edition of our text we have included various new topics and emphases throughout. These include the following:

- A person-centred approach to health research which generates evidence for patient-centred health care.
- New material on translational research which is concerned with how research-generated knowledge may be translated into effective new treatments and health services.
- New material on ‘mixed’ research methods.
- The inclusion of updated, contemporary published research for use in the revision questions.
- Cost issues in public health and primary health care.
- More discussion of clinical guidelines and systematic review methods.

We have included this new material because research methods, like all branches of knowledge, continue to improve and change. We are committed to ensuring that clinicians have ready and accessible training and

knowledge in these methods. Having a sound knowledge of contemporary clinical research methods enables you to rigorously evaluate evidence concerning clinical therapies. This is a foundation for the lifelong learning process that is required to deliver contemporary and effective clinical practice. Patients and the community expect contemporary knowledge to guide the recommendations of their clinicians and their treatment choices. To enable this process is a key purpose of our text.

We wish you all the best in your research training and the clinical practice it supports.

Steve Polgar, PhD

Shane Thomas, PhD

SECTION 1

Methodological Foundations of Health Research

OUTLINE

Introduction

1. Foundations of Health Research
2. Quantitative and Qualitative Methods
3. The Research Process

Introduction

In this section, we will examine the basic characteristics of the scientific method. A *method*, in the present context, refers to a system for acquiring knowledge and establishing its truth. Health providers frequently justify their theories and practices on the grounds that they are ‘scientific’, that is they are based on the use of scientific methodology. Scientific method is essential for conducting research and evaluation aimed at producing credible evidence, which is in turn used to improve the effectiveness and cost-effectiveness of health services.

In [Chapter 1](#), we discuss the fundamental principles for conducting health research and examine how these principles are applied to evidence-based practice. We examine the meaning of constructs such as facts, hypotheses, theories and paradigms and discuss how these are used in constructing knowledge.

In [Chapter 2](#), we discuss the elementary characteristics of both quantitative and qualitative methods of health research. The position taken in the present book is that the scientific method, as applied to health care, should preferably include both quantitative and qualitative methods. These two methods are used to produce the evidence for solving health-related problems. Although there are controversies regarding the nature of science and its methodology, there is a broad consensus concerning the principles and rules for producing scientific evidence. Our goal is to convey these principles and rules to our readers.

In [Chapter 3](#), we apply the basic methodological principles applicable to the conduct of patient-centred research outlined in [Chapters 1](#) and [2](#). This chapter outlines the stepwise problem-solving approach to research adopted as followed in this book. We will also relate the problem-solving approach to the structure of professional peer-reviewed publications reporting the results of health research.

Foundations of Health Research

Introduction,
Knowledge and Methods,
 Tradition,
 Reasoning,
The Scientific Method,
Positivist Paradigm,
 Observations, Description and Measurement,
 Generalization and Induction,
 Hypotheses and Laws,
 Theories,
 Deduction,
 Controlled Observation,
 Verification and Falsification,
 Research and the Advances of Health Care,
Postpositivist Paradigms,
 What Constitutes Falsification?,
 The Theory-Dependence of Observation,
 Science in a Social Context,
Pragmatism: The Combined Use of Quantitative and
Qualitative Methods,
Summary,

Introduction

Health research is a systematic and principled way of obtaining the evidence (data, information) for solving health care problems and investigating health issues. Research is *systematic* in that researchers follow a sequential process and *principled* in that research is generally carried out according to explicit rules. These rules or principles constitute the *method*. The term '*method*' derives from the Greek word '*methodoi*', meaning a 'way of travelling' or 'the path which we follow'. The application of appropriate methods guarantees the validity of the evidence and the truth of knowledge. In health research, *method* refers to a set of rules that specify the following:

- How knowledge should be acquired.
- The form in which knowledge should be stated.
- How the truth or validity of knowledge should be evaluated.
- How knowledge should be applied to advancing the safety and efficacy of practices.

Strictly speaking, the terms '*method*' and '*methodology*' have different meanings. Method refers to the rules of evidence followed for collecting data in a project. Methodology refers to the critical discussion, comparison and application of methods. The present chapter is concerned with methodology. Later chapters in this book are concerned with specific methods used in health research.

The primary aim of this chapter is to examine the philosophic foundations of research methods. Emphasis is placed on method as a means of conducting applied research and producing the best possible evidence for solving health problems.

The general aims of this chapter are to:

1. Examine the relationship between knowledge and methods.
2. Outline what constitutes the scientific method.
3. Examine positivist and postpositivist approaches to methodology.
4. Discuss the pragmatic position which is applicable to research and practice in contemporary health care.

Knowledge and Methods

Two concepts drawn from philosophy are relevant to our discussion: *ontology* and *epistemology*.

Ontology refers to the question of what exists in the world, what is ‘real’. Different knowledge systems take diverse positions on what constitutes reality. In contrast to the natural sciences, various traditional interpretations hold different views on what constitutes the reality. For example, the notion of life force or qi/chi, a central concept in traditional Chinese medicine, is absent in contemporary Western medicine. Of historical interest, is the classical Greek belief in humors that was once central to Western medicine. The balance of humors was thought to determine mental and physical wellbeing; however humors are no longer seen as ‘real’ in Western medicine.

Epistemology is a field of philosophy concerned with the nature, source and legitimacy of knowledge. In the domain of health research we are interested in knowledge as *applicable* to:

- Selecting and implementing practices.
- Producing and interpreting evidence.
- Constructing and applying theories to practice.

Before we begin discussion of scientific knowledge and research it is useful, as a means of contrast, to look at some other epistemological approaches.

Tradition

There are hundreds of different medical traditions. They are based on distinct ontological (the meaning of ‘being’) and epistemological (theory of knowledge) positions, ranging from shamanism to empirically oriented components of traditional Chinese medicine (TCM). The beginnings of TCM can be traced back thousands of years, well before the introduction of Western medicine. The principles of TCM were based, in part, on the experiences of practitioners, successfully treating diseases using natural methods such as acupuncture and herbal remedies. This practical

knowledge was combined with Chinese philosophy to formulate holistic theories of physiology, pathology and treatment. ¹ These theories emphasize the close relationship between the human body as a system integrated with the natural and social environments. The prevention or treatment of disease is contingent on the reestablishment of a dynamic balance between the energy flows within the body, the mind and the environment.

Reasoning

Reasoning is commonly used to arrive at true knowledge. It is assumed that, if the rules of logic are applied correctly, then the conclusions are guaranteed to be valid. As an example, let us look at the syllogism that follows:

1. All persons suffering from heart disease are males.
2. Person X has heart disease.
3. Therefore person X is a male.

The rules of logic guarantee that conclusion (3) is ‘true’, provided the syllogism is in a valid form and that premises (1) and (2) are valid. Modern science originates from ‘natural’ philosophy. Reasoning and logic are the foundations for scientific enquiry. However, we require valid evidence to support the conclusions we reach based on logic and the mathematical operations we perform on data to support the conclusions.

The Scientific Method

Science and the scientific method have evolved over a period of thousands of years. ² Great civilizations such as Babylonia, China and India devised written languages and symbols for numbers to permanently record observations and speculations about the world. This was an essential step in the development of formal science. The Greek and Roman philosophers and physicians further developed disciplines such as astronomy, mathematics and medicine. For example, the Roman physician Galen worked as a surgeon treating injured gladiators and used experimental methods to test hypotheses about physiology and anatomy. Much of the classical

knowledge was discarded during the Dark Ages, but an important fraction was preserved and expanded on by scholars from a wide variety of backgrounds, including Muslim, Chinese and Christian scholars.²

The beginnings of modern ‘Western science’ are generally traced to the beginning of the 16th century, a time in which Europe experienced profound social changes and a resurgence of great artists, thinkers and philosophers. Gradually, scholars’ interests shifted from theology and armchair speculation to systematically describing, explaining and attempting to control natural phenomena. These changed cultural circumstances enabled philosophers such as Descartes and Francis Bacon to challenge tenets of medieval thinking, and scientists such as Galileo, Newton and Harvey to propose new ways of conducting research and building knowledge. This period is referred to as the *Scientific Revolution*.³

The following points represent the essential characteristics of the early scientific worldview:

- *Realism*: the position which holds that the world exists independently of our beliefs. For example, we assume that the planets are large objects which circle about the sun, regardless of the observations astronomers make about their orbits. The work required of scientists is to discover and accurately describe the true nature of the world.
- *Determinism*: the assumption that events in the world occur according to regular laws and identifiable causes. The challenge for health and medical scientists is to identify the multiple factors which either cause or prevent illness and disease.
- *Empiricism*: the conviction that discovery ought to be conducted through observation and the truth of knowledge verified through evidence.
- *Scepticism*: an attitude which fosters questioning the truth of any proposition – even those made by great authorities. All aspects of knowledge, including the methods we rely on, are open to questioning, critique and revision.

These tenets remain an integral part of contemporary health and medical sciences. However, there are several interpretations of what constitutes ‘scientific method’. The fact of the matter is that there is no stone tablet on which the permanent axioms of the scientific method are inscribed. They are not eternal truths but rather conventions which have been found useful in advancing scientific knowledge and technology. The interpretation of what constitutes the scientific method is an on-going activity pursued by philosophers of science and research scientists.

Positivist Paradigm

The term ‘*paradigm*’ has multiple meanings depending on the context in which it is used. Here we are using it to refer to basic belief systems, which represent the overall methodological positions adopted by the researchers. Guba & Lincoln define a paradigm as:

a worldview, that defines, for its holder the individual’s place in it, and the range of possible relationships to that world and its parts, as, for example, cosmologies do. ⁴

Positivism is an influential interpretation of the science and its methodology. It was initially described by 19th-century philosopher August Comte. Currently, there are numerous versions of what is meant by positivism. [Fig. 1.1](#) represents our point of view.

Observations, Description and Measurement

Considering [Fig. 1.1](#), let us start with observations. The description of phenomena involving the precise, unbiased recording of observations of aspects of persons, objects and events forms the empirical basis of all branches of science. Observations can be expressed as either verbal descriptions or sets of measurements (see [Section IV](#)). The perceptions of the investigator must be transformed into descriptive statements and measurements that can be understood and replicated by other investigators. Some research is based on observations made with instruments (such as recording electrodes, microscopes and standardized clinical tests), whereas

other research calls for observation unaided by instruments. Although advances in instrumentation have contributed enormously to scientific knowledge, the use of complex instrumentation is not a necessary feature of scientific observation. Rather, the key attributes of scientific observation are accuracy and replicability by other scientists and practitioners. When observations are appropriately summarized and confirmed, they constitute the data, which is the factual bases of scientific knowledge.

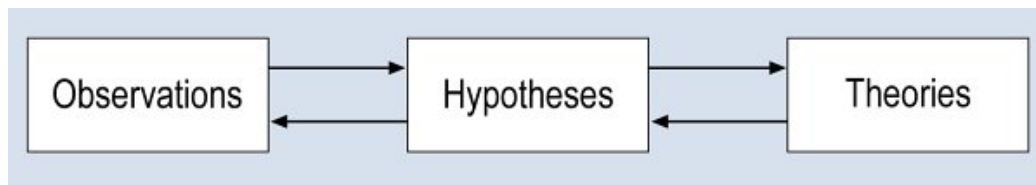


FIG. 1.1 The scientific method.

Generalization and Induction

Statements representing observations or measurements (data) are integrated into explanatory systems called *hypotheses* and *theories*. The logic underlying scientific generalization is called *induction*. Induction involves asserting general propositions (hypotheses, theories) on the basis of a limited number of observations of select elements. For example, having observed that penicillin is useful for curing pneumonia in a limited set of patients, we make the generalization that ‘the administration of penicillin cures pneumonia (in all patients)’. Generalizations or inferences are seen as probabilistic. How probabilistic inferences work will be discussed in later chapters.

Hypotheses and Laws

‘The administration of penicillin cures pneumonia’ is an example of an hypothesis. Scientific hypotheses are statements that specify the expected relationship between two or more sets of variables. Variables are specific, observable and measurable properties of persons, objects or events. In this instance the first variable relates to the administration of penicillin and the

second to beneficial changes in patients with pneumonia. As seen in subsequent sections, an important feature of scientific hypotheses is that the terms used must have clear-cut, observable referents. For example, the ‘beneficial changes in the symptoms of patients with pneumonia’ will include variables such as ‘temperature in degrees centigrade’. When hypotheses acquire strong empirical support, they may be called laws. Therefore the statement, ‘the administration of penicillin cures pneumonia’, may be considered a law on the grounds that many patients with pneumonia have been effectively treated. Nevertheless, there are problems concerning the universal truth of this statement (see the section ‘Verification and Falsification’). We will look at this issue in a practical context, under sampling (see [Ch. 5](#)), using samples for statistical inference.

Theories

Scientific theories are essentially conjectures representing the current state of knowledge about the world. Hypotheses or laws are integrated into more general explanatory systems called *theories*. A theory will clarify the relationships between a network of observations and hypotheses. For example, a theory to explain why drugs called antibiotics are effective in curing some infectious diseases integrates evidence from diverse sources such as microbiology, pharmacology, cell physiology and clinical medicine. It is an essential feature of scientific theories that they are based on the correct use of language and logic. Some theories entail a model, which is a mathematical or physical representation of how the theory works. A simple example is an anatomic model of the circulatory system. In this way, theories explain the causes of events and provide conceptual means for predicting and influencing relevant events. Theories are important for explaining the causes of health and illness and predicting the probable effectiveness of treatment outcomes. For example, the theory of AIDS is based on the assumptions that signs and symptoms of the disease are caused by the HIV retrovirus, which attacks the human immune system. Furthermore, it is assumed that the HIV virus is transmitted by bodily fluids (e.g. blood, semen). This theory guides a variety of effective practices such as screening of blood and promoting safe sex to control the pandemic.

Deduction

A scientific theory leads to a set of empirically verifiable statements or hypotheses. In addition to being generalizations based on evidence, hypotheses also may be deduced logically from the theories and/or models which specify the causal relationships postulated by the theories. For instance, if we hold the theory that the patterns observed in the activity of a set of neurons in the occipital lobe mediate visual sensation in humans, then the hypothesis follows that the activation of these neurons (say, by electrical stimulation) will lead to the report of visual sensations. Such hypotheses have been the bases for subsequent spectacular clinical advances, such as artificial vision through cortically implanted electrodes.

Controlled Observation

Whenever possible, it is desirable that hypotheses should be tested under controlled conditions. The aim of control is to discount other competing hypotheses for explaining the predicted outcome. For example, if we intend to show that occipital lobe stimulation causes the visual sensations, then we must control for other types of brain stimulation causing this outcome. Conversely, we would need to show that occipital lobe stimulation does not lead to a host of other types of sensations. Only by discounting alternative explanations through control can we have confidence in the relevance of our evidence relating to our research hypothesis.

Verification and Falsification

As discussed previously, after the data have been collected, the investigator decides whether the findings are consistent with the predictions of the hypothesis. If the evidence supports the hypothesis, then the theory from which the hypothesis was deduced is strengthened or verified. When the data are not consistent with the predictions, the theories from which the hypotheses were deduced are said to be falsified. If a theory can no longer predict or explain the evidence in its empirical domain, it becomes less useful and is usually discarded in favour of new, more powerful theories. Therefore scientific theories are not held to be absolute truths but rather as

probabilistic explanations of available evidence. An essential characteristic of a scientific proposition hypothesis is that it should be ‘falsifiable’. That is, there should be a clear empirical outcome that could demonstrate that the proposition was false. For example, consider the previous statement, ‘the administration of penicillin cures pneumonia’. This statement is clearly falsifiable because there is the possibility that penicillin will not work as expected. If a proposition cannot be falsified then it is not a scientific proposition.

Research and the Advances of Health Care

The application of the positivist paradigm to investigating the world has contributed to the spectacular growth of scientific knowledge. Research relying on observation and measurement, facilitated by new instrumentation, resulted in the discovery of an enormous number of accurate and reproducible facts about health and illness. New facts both challenge existing theories and call for the creation of novel, more powerful theories. The new theories serve as impetus for more research, resulting in new instrumentation and observations. Advances in scientific knowledge have been applied to creating new technologies, which in turn contribute to new discoveries and advances in scientific knowledge. For example, the invention of computers was possible because of advances in electronics, chemistry and mathematics. In turn, the use of computers has immensely facilitated obtaining and summarizing scientific observations and generating explanatory models. Additionally, the use of computers as information-processing systems has generated useful metaphors for theoretical advances, such as explaining the workings of the human brain and mental functioning. In this way, the scientific method contributes to advancing theory and practice, helping us to describe, explain, predict and control the world in which we live.

Postpositivist Paradigms

The description presented earlier is not the only interpretation of the scientific method. Rather, there are compelling reasons why the positivist view was challenged in the 20th century. Advances in science contributed to

changes to our methods of conducting enquiry and the ways in which we explain the world. For example, ‘quantum theory’ and the ‘theory of relativity’ challenged the mechanistic, clockwork-like view of reality. Modern theories of physical reality are ‘fuzzier’ and more probabilistic than the previous, more ‘clockwork’ like positivist views. The observers are not seen as being outside the phenomena they are studying. The following discussion addresses some of the ideas of the postpositivist approach to science.

What Constitutes Falsification?

It was stated earlier that when new credible and valid evidence is inconsistent with the predictions of a theory, the theory is ‘falsified’ and is eventually modified or discarded. Commentators argued that theories are not, in practice, so readily modified or discarded by the scientists who are committed to them. Rather, they are structures which have an inner ‘hard core’ of propositions protected by an ‘outer belt’ of auxiliary, modifiable hypotheses. Evidence inconsistent with predictions based on the theory results in the modification of auxiliary hypotheses, rather than discarding the ‘core’. Consider, for example, the ‘germ’ theory of infectious disease, on the basis of which one would predict that penicillin (which kills ‘germs’) will cure bacterial infections. Suppose that we administer penicillin to a number of patients with an infectious disease and find, contrary to what was predicted, no clinically useful changes are observed. On the basis of this falsification, would we discard the germ theory of disease? Probably not. Rather, we will use an auxiliary hypothesis to explain our findings, such as ‘the development of penicillin-resistant bacteria’. Of course, eventually we would need to demonstrate that such bacteria do, in fact, exist. If no clear evidence is available to support the auxiliary hypothesis, then it would be dismissed as a credible alternative explanation.

The methodological issue which remains controversial is the logical basis for discarding one theory and accepting its rival. Researchers judge the outcome of a research program in the context of not only particular finding but rather by an overall pattern of related findings and theories.

The Theory-Dependence of Observation

Critics argued that it is simplistic to believe that observations are made independently of theoretical notions held by the observer. The observer is selective with regard to what is recorded as data and then subsequently analysed as evidence. Our observations and facts are ‘theory-dosed’; that is, theories specify which observations are of importance and what aspects of these observations should be recorded or ignored. For instance, in observing the electroencephalogram (EEG) of a patient with epilepsy, our perception is guided by theories of the electrical activity of the brain and the nature of brain pathology. We will also hold ideas about how the EEG machine works (e.g. electrode sensitivity, amplification, etc.), and there are rules to identify artefacts in the evidence. What is observed as evidence of epilepsy by an expert could well be perceived as meaningless squiggles by a naive observer.

Also, we can assume that human participants are actively engaged in the research process. Criteria for conducting ethical research (see [Ch. 6](#)) require that participants be informed of the aims, risks and potential benefits of a research project. It follows that participants will have expectations regarding the processes and outcomes of the project. These expectations may be a source of bias and influence the facts. There are various ways for controlling for bias, as discussed in later chapters.

Science in a Social Context

Health research is conducted in particular social settings by individuals with personal aims, values and social settings. There are multiple perspectives from which researchers may view a given problem. For example, professional affiliation is a particularly important factor in deciding which research questions are investigated and how the evidence is to be applied. Consider the example of the growing problem of obesity in developed countries. A public health researcher would look at programs which would contribute to preventing/reducing obesity in a given population. Certain programs might be devised and evaluated, such as:

- Mandatory labelling of manufactured food (e.g. kilojoules, trans fat, sodium, sugar, etc.).
- Promotion of healthy eating (e.g. restrictions on the type of food sold in school canteens).
- Government-funded promotion of healthy living programs (e.g. tax incentives for joining gyms, community-based sporting events, etc.).

In contrast, a medical researcher would focus on the efficacy of therapeutic interventions, such as:

- Medically supervised diets and exercise for individual patients.
- Medication which suppresses appetite or inhibits the absorption of fats.
- Surgery for morbidly obese patients (e.g. gastric banding).

Until recently, the ‘medical model’, was the most dominant approach to identifying and solving health problems. The role of the health professional was to identify the underlying pathology and to implement appropriate treatment to correct the problem. The patient is assigned a rather passive role in this process, being the ‘locus’ of the problem, and is expected to comply with medical advice. The aim of medical research is to enhance the range and efficacy of treatments which can be offered to patients.

The holistic approach, which also informs health care research, is referred to as the *biopsychosocial* model.⁵ The biopsychosocial model currently has considerable influence in how health care is conceptualized and research is conducted. At the same time, the medical model, strongly based on positivist thinking, remains influential in research aimed at understanding the workings of the human body and improving the technical aspects of health care. A postpositivist paradigm retains the realist ontology of positivism but emphasizes that reality is interpreted from different perspectives. These perspectives are essential and are best researched using qualitative methods.

Pragmatism: The Combined use of Quantitative and Qualitative Methods

A postpositivist paradigm recognizes different perspectives of reality. In this book, we will explore two of these approaches to conducting health research: *quantitative* and *qualitative*. Using a quantitative approach, we view our patients objectively, as natural objects, and attempt to identify and measure important variables which represent the causes and expressions of health problems. We develop models and theoretical frameworks which systematically explain how these variables are interrelated and undertake therapeutic actions which serve to diminish the values of the variables representing illness or disability. Our therapeutic actions are the technical applications for our scientific theories; their outcomes and effectiveness should be tested under controlled conditions. Using a qualitative approach to research and knowledge, we view our patients as individuals and attempt to gain insights into their subjective experiences and the reasons for their actions. We develop theories for interpreting personal points of view and to inform our therapeutic actions so that they are meaningful and appropriate to our patients.

Not all scholars agree that qualitative and quantitative methods can be logically combined. It has been argued that these two methods are incommensurate in that an *objective* view of reality need not necessarily correspond with that of a *subjective* experience.⁶ The controversy was resolved on pragmatic grounds, based on the argument that both methods are needed to solve problems in applied areas such as health care.^{6,7}

Pragmatism is a system of philosophy that avoids speculation about nebulous abstractions such as ‘The Truth’. Rather it defines truth as what works to enable us to provide the best possible health care for our patients and the community. A pragmatic, problem-solving approach is followed in the present book, emphasizing the need for both qualitative and quantitative methods for effectively solving health problems. The introduction of a biopsychosocial approach has raised questions about the combination of methods relevant to health sciences research. We perceive patients or clients in two different but interrelated frameworks: first, as broken down or malfunctioning biological systems; and second, as persons, like ourselves,

living in a society and attempting to make sense of and cope with their particular health care problems.

Summary

Methods are used for acquiring, stating and establishing the truth of knowledge. The scientific method underpins the validity of theory and practice in Western health care. The scientific method is concerned with applying a set of rules or conventions that will allow us to produce scientifically valid knowledge. These rules specify how observations should be made, and how theories and hypotheses should be stated and evaluated.

Theories and hypotheses obtained and verified through scientific enquiry are not held to be absolutely 'true'. An inherent part of the scientific approach is scepticism regarding both the contents of knowledge and the underlying methodology. We have pointed out that there are controversies concerning what constitutes scientific methodology. It was argued that the scientific method is directly applicable to conducting research in the health sciences. We have adopted a postpositivist paradigm which entails the use of both qualitative and quantitative methods. A patient-centred approach to research takes into account both objective facts associated with health and illness and the patients' beliefs and interpretations and responses to these.

References

1. Lu A.P, Jia H.W, Xiao C, Lu Q.P. Theory of traditional Chinese medicine and therapeutic method of diseases. *World J Gastroenterol* . 2004;10(13):1854–1856.
2. Fara P. *Science: A Four Thousand Year History* . New York, NY: Oxford University Press; 2009.
3. Okhasa S. *Philosophy of science: a very short introduction* . 3rd ed. Oxford University Press; 2016.
4. Guba E.G, Lincoln Y.S. Competing paradigms in qualitative research. In: Denzin N.K, Lincoln Y.S, eds. *Handbook of Qualitative Research* . London: Sage; 1994.

5. Engel G.L. The need for a new medical model: a challenge for biomedicine. *Science* . 1977;196:129–136.
6. Minichiello V, Aroni R, Hays T. *In-Depth Interviewing* . 3rd ed. Sydney: Pearson Education Australia; 2008.
7. McGartland M, Polgar S. Paradigm collapse in psychology: the necessity for a ‘two methods’ approach. *Aust Psychol* . 1994;29(1):21–28.

Quantitative and Qualitative Methods

Introduction,
The Social Construction of Reality,
Contrasting Qualitative and Quantitative Methods,
Perception of Subject Matter,
Positioning of Researcher,
Data and Evidence,
Theories,
Theory Testing,
Applications in Health Care Delivery,
Summary,

Introduction

In [Chapter 1](#), we discussed the conceptual foundations of health research methods. We suggested that it was advantageous to conduct health research within a pragmatic framework that reflects real health problems and treatments. The pragmatic approach includes both quantitative and qualitative methods. In terms of the quantitative approach, health problems are frequently conceptualized as the impairments and dysfunctions of an organism, a machine that is malfunctioning. The aim for researchers is to

accurately describe these problems and identify their causes and consequences. Accurate, evidence-based knowledge enables practitioners to prevent illness and mitigate effects of the associated impairment and dysfunction. This approach to research and practice conceptualizes patients as if they are broken mechanisms in need of repair. However, ignoring personal values and experiences may significantly diminish the effectiveness of health interventions. The human machine is not just a machine; it is, at the least, a thinking machine. And, according to many religious frameworks, it is a thinking machine with a soul. Qualitative research provides an important perspective of the nonmachine elements of all of the participants in clinical settings. There is strong evidence that clinical outcomes are directly affected by patient experience. Doyle, Lennox and Bell's 2013 review of 55 well-designed studies of this relationship strongly supports this conclusion. Their conclusion that 'Clinicians should resist sidelining patient experience as too subjective or mood-oriented, divorced from the "real" clinical work of measuring safety and effectiveness' has been confirmed in many studies and is now considered to be a key component of many clinical guidelines. ^{1,2} Qualitative research gives a voice to research participants, allowing researchers to gain valuable insights into how health problems are viewed and experienced by the research participants. These insights are clinically valuable and affect clinical outcomes. In this chapter, we will examine and contrast these two methods in greater detail.

The aims of this chapter are to:

1. Outline different conceptual approaches to health research.
2. Compare and contrast specific dimensions of qualitative and quantitative approaches to advancing knowledge and effective practice.
3. Emphasize the importance of using evidence from both qualitative and quantitative perspectives.

The Social Construction of Reality

As discussed in [Chapter 1](#), knowledge is acquired and applied from a particular perspective. An example outside health research will illustrate the

contrast between qualitative and quantitative perspectives. Imagine that you are given a bank note and it has $\leq \$100$ printed on it. You can examine this object from a number of perspectives depending on who you are and how you are positioned in relation to this object.

- If you were a person from a remote part of the world with no exposure to such bank notes, you might see it as a decorative object and value it as a curio. (Qualitative.)
- If you were a person having financial difficulties, $\leq \$100$ would be seen as a means to buy food or cover bills. (Quantitative.)
- If you were an accountant, you would focus on the number printed on the bill for making fiscal calculations. (Quantitative.)
- If you were a counterfeiter (we hope not!), you would be intensely interested in the physical properties of the object, such as its length, width, colour and texture: details required for successful forgery. (Quantitative.)

Money is very much a social construct. It is 'real' enough; having it or not having it can influence all aspects of life, including the level of health services a person can access. At the same time, money is an abstraction; it can take virtually any physical form (e.g. coins, notes, credit cards or a set of numbers representing your bank balance). The reality of money is constructed through the actions of a number of institutions (e.g. banks, state treasuries, etc.) which determine its value for buying goods and services. The point here is that health researchers can look at a problem from different perspectives and select the appropriate method to solve it depending on the questions they are asking and the nature of the information they wish to acquire.

The importance of integrated qualitative and quantitative research approaches is reflected in the strong emphasis that is now given to patient-/person-centred care in contemporary health systems and services. There are many policy and scholarly papers about patient-centred care; however, a good definition is offered by the Institute of Medicine: 'Providing care that is respectful of, and responsive to, individual patient preferences, needs and values, and ensuring that patient values guide all clinical decisions'. Various international health agencies such as the NHS

(United Kingdom: <https://www.england.nhs.uk/integrated-care-pioneers/resources/patient-care/>) also provide extensive discussions about the desirability of patient-centred care in contemporary health care. This contemporary emphasis requires that health research methods can accommodate the rigorous study of patient values, preferences and needs, and hence we are obligated to use both qualitative and quantitative approaches.

Contrasting Qualitative and Quantitative Methods

In this section, we will identify and discuss the fundamental differences between quantitative and qualitative methods (Table 2.1). One key difference is that these two methods have different historical and disciplinary backgrounds. Quantitative methods are based on the traditions of the natural sciences (physics, biology, etc.), whereas qualitative methods emerged from disciplines such as philosophy, anthropology and other social sciences (see Ch. 09).

Perception of Subject Matter

As discussed in Chapter 1, quantitative researchers favour a ‘naive’ realistic view of knowledge. The associated research strategies focus on the discovery of novel facts about the world. *Reductionism* is central to conducting quantitative research. Reductionism refers to the process of reducing complex phenomena to simpler, more fundamental elements. For example, the advancement of the discipline of chemistry has been greatly enhanced by reducing the immense variety of materials to the interaction and combination of elements, represented in the periodic table. Also, the enormous array of physical and mental symptoms has been compiled into taxonomies of illnesses. Good health can be differentiated from ill health along with a number of clearly defined factors, such as blood pressure, cholesterol levels or the presence or absence of infectious agents. For example, ischaemic heart disease can be attributed to diminished blood flow

in the coronary arteries. This in turn can be linked to risk factors such as high cholesterol or high blood pressure.

TABLE 2.1**Contrast between Quantitative and Qualitative Methods**

	Quantitative	Qualitative
Perception of subject matter	Reductionistic: identification and operational definition of specific variables	Holistic: persons in the context of their social environments
Positioning of researcher	Objective: detached observation and precise measurement of variables	Subjective: close personal interaction with participants
Database	Quantitative: interrelationships among specific variables	Qualitative: descriptions of actions and related personal meanings in context
Theories	Normative: general propositions explaining causal relationships among variables	Interpretive: providing insights into the nature and social contexts of personal meanings
Theory testing	Controlled: empirically supporting or falsifying hypotheses deduced from theories	Consensual: matching researcher's interpretations with those of participants and other observers
Applications	Prediction and control of health-related factors in applied settings	Interacting with persons in a consensual, value-consonant fashion in health care settings

Adapted from McGartland & Polgar (1994). Copyright (1994) The Australian Psychological Society Ltd. Reproduced by permission.

Understanding the risk factors and causes of heart disease enables the implementation of interventions to reduce the burden of this condition on individuals and in the general community. To identify causal associations, clearly stated hypotheses need to be tested in research studies (see [Ch. 1](#)). An important example of this is the relationship between factors or variables representing changes in signs and symptoms of a disease. For example, Lipitor is effective in lowering blood cholesterol, thus reducing the incidence of heart attack.

However, by reducing health problems to specific disease processes and associated signs and symptoms, quantitative researchers may lose track of the individual with the health problem. A *holistic* patient-centred perspective focuses on the person with the medical problem in the context of the individual's physical and social environment and their values, needs and preferences. Qualitative researchers aim to understand how individuals experience health problems and the reasons for their actions to cope with these problems. For example, qualitative researchers investigating heart disease would consider the participants' experiences and understanding of their health issue. Therefore the person at risk of a heart attack is no longer reduced to a failing mechanism, but rather is perceived as an active agent for achieving positive changes.

Positioning of Researcher

Quantitative research is conducted to discover valid and reliable facts about the world, as discussed in [Chapter 1](#). The fundamental positioning of the researcher is '*objective*', aiming to perceive and record events dispassionately, without any personal bias or distortion. However, there is recognition of the possibility of error and bias entailed in collecting data. Therefore quantitative health research has invested much effort in devising and evaluating accurate measuring instruments and tests (see [Ch. 14](#)). Also, research designs such as 'double-blind randomized controlled trials' (see

Ch. 7) have been implemented to reduce the effects of both observer and participant bias.

The problem with highly structured data collection is that participants' true responses may be restricted or hidden because of the method employed to uncover them. Qualitative researchers find standardized instruments intrusive, as they may restrict self-expression and they may impede the understanding of the actual experiences, ideas and emotional responses of the participants. To understand personal meanings and subjective experiences, one has to become involved with the lives of the participants being studied. In this way, the researcher takes a subjective position, a degree of empathy may develop between the researcher and the participant and this may enhance consideration of the patient's values, needs and preferences as required by patient-centred care principles. By empathy we mean the ability to 'put ourselves in the other person's shoes', that is, to see things from their perspective(s).

Data and Evidence

By 'data' we mean the results of a study which are collected and analysed to provide the evidence required to solve a research problem. The data obtained in quantitative research consist of sets of observations and measurements, descriptions of physical and behavioural events, health and illness. These are summarized and analysed in accordance with statistical principles outlined at an introductory level in Sections V and VI.

In qualitative research, the database is essentially an analysis of the narrative (or a story, if you like) that reports what has happened to people, what they did or said in specific situations. This narrative should be adequately detailed so as to illuminate for the reader the personal meanings that the health-related events had for the informants.

The data in qualitative research are descriptive, a detailed or thorough description of what people said, their actions and activities, and interactions with other people. An important aspect of qualitative research is keeping thorough, up-to-date field notes. These should be recorded as closely as possible to the time of occurrence of the phenomena under study. The field notes should contain direct quotations from the participants and the settings in which the statements and actions were recorded. Where possible (where

it is appropriate and not overly intrusive), the researcher may use audio and video recordings. This helps record interviews, and it improves accuracy in conveying what was said and done in a given setting, since it is possible to review the obtained information.

Theories

Theories represent our current state of knowledge about the state of the world. Theories are abstract, coherent explanatory systems that integrate and accommodate a broad range of research findings. Theories may be constituted of premises stated in everyday language, with particular attention paid to the appropriate use of concepts and the logical development of the premises.

Theories based on quantitative evidence integrate patterns of findings concerning the interrelationships among variables. Such theories often contain ‘models’, which may be mathematical and/or systems representations of the patterns of findings. Models of anatomic and physiologic processes, such as those of the circulatory or nervous systems, are good examples of successful quantitative models. Quantitative theories are ‘normative’ in the sense that they aim to describe and explain, as closely as possible, how things ideally work. They represent what is assumed to be stable and universally applicable. Conversely, theories integrating evidence from qualitative research do not address facts about how objects are constituted and interact, but rather are the overall interpretations of personal meanings emerging in specific social settings.

Qualitative health researchers approach theory formation by attempting to identify common ‘themes’ or categories of meanings emerging from the data. The important point here is that the theoretical categories are developed from evidence expressing personal meanings, rather than ‘facts’ derived from the statistical treatment of objective measurements of specific variables. In this way, theory is said to be ‘grounded’ in the narratives of a group of individuals.

Theory Testing

Theories based on quantitative evidence are tested through precisely stated predictions or hypotheses logically deduced from the theories (see [Ch. 1](#)). The accuracy of a quantitative theory is judged by the extent to which the predictions generated by the theory match the evidence produced by methodologically rigorous research. However, as discussed in [Chapter 1](#), theories are not easily discarded on the basis of preliminary, insufficient evidence.

Testing qualitative theories is somewhat different, as no causal mechanisms are included in the theoretical framework. The simplest verification of qualitative interpretations is to go to the participants themselves to establish if the researcher's interpretations make sense to them. The extent to which a consensus develops between researchers and their participants is one of the important indications of the truth of qualitative theories.

Applications in Health Care Delivery

The applications of quantitative evidence and theories are essentially technical, providing mechanisms in terms of which we can *predict* and *control* specific health-related variables. That is, we apply quantitative approaches for (1) discovering the causes of diseases and disabilities, (2) predicting the burden of diseases and disabilities on individuals or populations, (3) developing and validating assessment procedures and (4) evaluating the effectiveness of interventions.

In contrast, qualitative research provides evidence and theories that enable us to better understand our participants as human beings. This research discloses how illnesses, disability and health care delivery affect people's lives interpreted from their points of view.

Summary

Qualitative research is based on the assumption that reality is *socially constructed*. It entails disciplined enquiry examining the personal meanings of individuals' experiences and actions in the context of their social environments. By *qualitative* we mean that the data consists of detailed descriptions based on language or pictures recorded by the investigator. The

term ‘*discipline*’ indicates that the enquiry is guided by explicit methodological principles for defining problems, collecting and analysing the evidence, and formulating and evaluating theories. *Personal meaning* refers to the way in which individuals subjectively perceive and explain their experiences, actions and social environments.

In contrast, quantitative research holds the view that reality is independent of personal views and social contexts. The role of the researcher is to discover the objects and processes as they exist in the world. Quantitative research also involves disciplined enquiry based on reducing issues and problems to defined variables. Data collection involves measurement and observation under controlled conditions, enabling researchers to generate mechanistic or systems theories for explaining how variables are interrelated. The data are presented in a numerical form and are analysed using statistical techniques.

Different research designs may be used to generate evidence of the same processes, although from different perspectives. For instance, any complex clinical phenomenon, such as schizophrenia, may be studied using any of the research strategies outlined in this text. To understand the scope of the problems and the effectiveness of the appropriate treatments, it is desirable to use a variety of research strategies. Conversely, a comprehensive theory of a clinical problem should generate any number of hypotheses within the realm of the research strategies discussed in this book. We examine the analysis of qualitative data in Chapter X.

References

1. Doyle C, Lennox L, Bell D.A. Systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* . 2013;3(1):e001570. doi: [10.1136/bmjopen-2012-001570](https://doi.org/10.1136/bmjopen-2012-001570).
2. Roddis J.K, Liversedge H.L, Ryder I, Woodhouse M. Incorporating the patient experience into clinical guidelines: recommendations for researchers and guideline

developers. *BMJ Evid Based Med* . Published Online:
September 18, 2018 doi: [10.1136/bmjebm-2018-111015](https://doi.org/10.1136/bmjebm-2018-111015).

The Research Process

Introduction,

Sequential Steps of the Research Process,

1. Research Problems and Questions,
2. Research Planning,
3. Design,
4. Data Collection,
5. Organization and Presentation of the Data,
6. Data Analysis,
7. Interpretation of the Evidence,
8. Evaluation and Dissemination of the Results,

The Structure of a Research Paper,

Title and Abstract,
Introduction,
Method,
Results,
Discussion,
References and Appendices,

The Publication Process,

Research and Evidence-Based Health Care,

Summary,

Introduction

The previous two chapters were concerned with conceptual and philosophic foundations of the scientific method. In this chapter, we will examine the method in a more practical way, as it is applied to collecting data and generating research evidence. Health and medical research is conducted as a stepwise, problem-solving approach. The steps for conducting research are reflected in the way in which research is reported in professional journals (which you will be reading throughout your career as part of the process of keeping up-to-date). Almost all clinical disciplines now have on-going mandatory requirements for continuing professional development (CPD), also known as continuing medical education (CME), that involve reading and understanding published health and medical research.¹ This is a fundamental component of quality improvement activity to ensure that clinicians have a contemporary knowledge of the best clinical practices. The central aim of this chapter is to enable you to understand the rationale for conducting original research, which is the foundation for your clinical knowledge and clinical practice. The specific aims of the chapter are to:

1. Outline the steps of the research process.
2. Describe the conventions for presenting published research.
3. Briefly outline the ways in which research publications are essential for evidence-based practice.

Sequential Steps of the Research Process

Although there are differences in how health scientists approach specific problems, [Fig. 3.1](#) shows the steps commonly followed in applied health research. You might recognize this as an ‘algorithm’, a general programmed approach to solving problems. To illustrate the sequential steps, we will refer to a published research paper titled ‘A qualitative study of GPs’ views of treating obesity’ by Epstein & Ogden.² You can access this publication from the *British Journal of General Practice* to enable you to follow the steps described here in more detail.

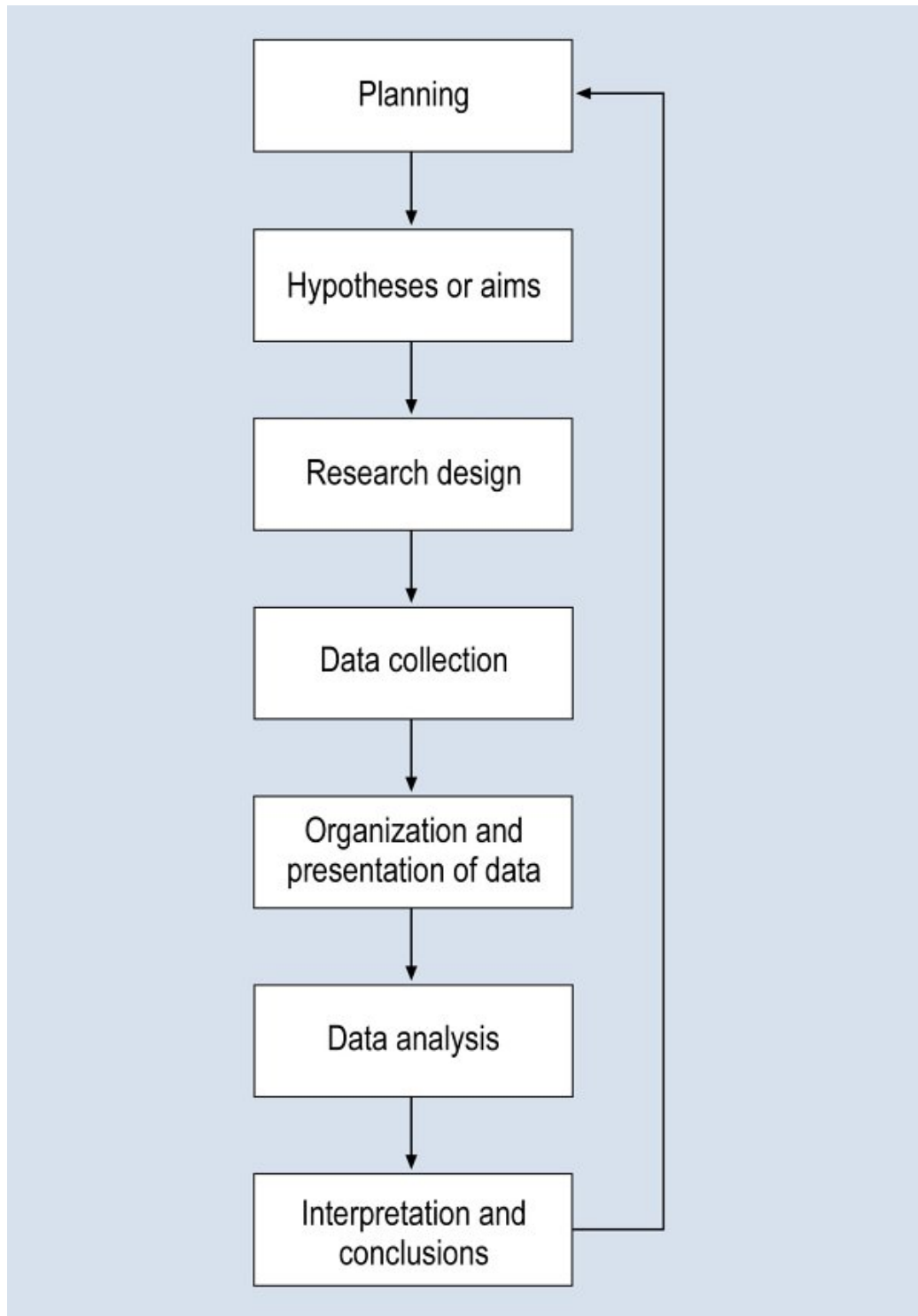


FIG. 3.1 The research process.

1. Research Problems and Questions

The first step in the problem-solving process is to state clearly and unambiguously the problem that you are intending to solve. Research problems reflect both the unmet health care needs of patients and the opportunities created by new theoretical and technological advances for improving health care (see [Chs 4 and 24](#)). The problem must be realistic, one that can be solved with the resources at the researchers' disposal. It is pointless and unethical to initiate projects that cannot be completed (see [Ch. 6](#)).

Having identified a problem, the next step is to ask the 'right' answerable research question. A well-formulated research question will guide the research project to produce the evidence required to answer the question and (if all goes well) to contribute to the solution of the research problem. [Chapter 4](#) will discuss in more detail the way researchers formulate questions.

In the case of the paper by Epstein & Ogden ², the research problem was identified as the lack of evidence concerning general practitioners' (GPs') views of their treatment approaches to obesity. The research question was: 'How do general practitioners view the treatment of obesity?' Therefore the aim of the researchers was to obtain evidence to describe and understand GPs' views using a qualitative approach. It is widely recognized that clinicians have a pivotal role in promoting the health of their patients and preventing the onset of serious avoidable diseases. ³

2. Research Planning

Note that the researchers may have used a structured quantitative method to obtain similar, pertinent evidence to address the research question. Of course, one may use both questionnaires and interviews in the same study. Obesity in the community is a serious global health problem. ⁴ It is understood that the research problem should lead to evidence which will improve practices and health outcomes.

Research planning (see [Section II](#)) involves selecting appropriate strategies and data collection techniques to answer research questions and to test the research hypotheses. (We explain what these are in [Chapter 4](#).) Research planning relies on detailed knowledge of previous research

summarized in a literature review. Also, the plan should take into consideration ethical and economic factors before the appropriate data collection strategies are implemented and the precise research aims are stated. Additionally, the planning process will take into account the target population and formulation of a sampling strategy to select the required sample.

In the present example, the population was defined as GPs working in London and the sample consisted of 21 GPs who consented to being interviewed for the study. The ethical requirements for this study were minimal, as there were no patients involved and the participants were commenting on their everyday professional experiences relevant to the research question.

3. Design

Research designs are clear statements of how the data are to be collected in the study. Appropriate research designs will lead to obtaining the data suitable for answering the research question (see [Ch. 7–10](#)). In the study we are discussing, the design is described as *qualitative* research using semi-structured interviews. This design was appropriate for obtaining data relevant to describing how GPs view the treatment of obesity.

4. Data Collection

The next step in the research process is collection of data. We will examine data collection methods employed in health research, including measurement, interviews and focus groups. In this study, the design that is the plan for data collection depended on semi-structured interviews with the doctors.

5. Organization and Presentation of the Data

We will examine basic concepts in this area, outlining how graphs and various descriptive statistics are used to condense and communicate research and clinical findings for qualitative research designs. When employing qualitative research, the results often are presented in the form

of direct quotations of what was expressed. In the present study, the researchers provided a series of direct quotations representing GPs' understanding and management of their patients' weight problems. For example, 'Dr 15' (Doctor 15) described an unhealthy diet of a patient as: 'Because of his lifestyle – he was working in the city – his lunch was fast food and so it was a difficult issue for me'.

TABLE 3.1**Format of Research Publications and the Research Process**

Publication format	Research process
Title	Explains the focus and content of the study
Abstract	Provides a brief informative outline of what was done and what was found
Introduction	Research planning. Answers the question, “why was this work necessary?”
Method	Design. Explains how the research was conducted
Participants	Explains who participated in the study
Apparatus	Measurement and data collection
Procedure	Provides a step-by-step description of how the research was conducted
Results	Descriptive statistics or narratives and inferential statistics or thematic analysis describe the study’s outcomes
Discussion	Interpretation of the evidence and recommendations for clinical actions and future research
References	A list of the work used to inform the study design and results
Appendices	Provided if necessary to explain the study and its results

6. Data Analysis

The analysis of quantitative data involves applying the principles of probability for calculating confidence intervals and testing the research hypotheses. Inferential statistics and decision making is outlined in [Chapters 18 to 19](#). The presentation and analysis of qualitative data involves identifying key themes which provide insights into participants' personal meanings of events and experiences. The way in which GPs conceptualized their responsibility in the management of obese patients and how they saw the role of the patients in this process were the two key themes which emerged from the interviews.

7. Interpretation of the Evidence

The next step in a research project is the interpretation of results (see [Ch. 21](#)). The evidence is used to answer the research question and may either support or falsify existing theories or practices. Also, for various reasons, a research project may fail to produce sufficient evidence for confidently answering the research question. It is rare that the findings from any single research project are completely definitive, and often the results may suggest the need for further investigation in related subject areas or contexts. In the discussion section, the authors reported that the GPs conceptualized obesity in terms of the patients' responsibility. It was suggested that future research was required to resolve an apparent conflict between the understanding of GPs' and patients' views of the management of obesity. Further research is required to provide evidence for improving communication between GPs and their patients.

8. Evaluation and Dissemination of the Results

For research to be meaningful, investigators must publish their results in professional journals, present them at conferences or both. Research findings become part of scientific knowledge only if the studies stand up to methodological critique and replication. We will outline steps for critically evaluating published research. Also, we note that evidence from various related publications can be synthesized for producing systematic reviews

guiding evidence-based practices in [Chapter 23](#). Epstein and Ogden disseminated their results by publishing their findings in the *British Journal of General Practice*. In doing so, the results of the study may generate further discussion and research regarding the role of GPs in helping people manage weight problems. At the time of writing, the paper was cited by 208 other researchers, so it had a significant effect on thinking in this area. The research effect of papers can be measured by the various databases and tools designed for this purpose, such as Google Scholar, Research Gate and Thompson citation tools. The translation effect of such studies also can be measured by the effect on clinical practice and clinical standards and guidelines.

The Structure of a Research Paper

The format of a professional publication reporting empirical studies generally reflects the stages of the research process discussed in this book. [Table 3.1](#) represents the relationship between the stages of research and the commonly used publication format for empirical studies. This form is not necessarily followed for all types of scholarly communications, such as for some qualitative studies, theoretical papers and literature reviews. In the subsections that follow, we examine in more detail each of the components of a research report shown in [Table 3.1](#).

Title and Abstract

The *title* is a descriptive sentence stating the exact topic of the report. Many titles of quantitative research reports take one of two forms:

- y as a function of x
- the effect of x upon y

In causal research, such as experiments, y refers to the dependent variable being measured and x refers to the independent variable being manipulated (see [Ch. 7](#)). For example:

- The incidence of alcoholism in health professionals as a function of work-related stress
- The effect of major tranquillizers on the cognitive functioning of persons with schizophrenia

For descriptive quantitative or qualitative research, the title should inform the reader about the groups being studied and the characteristics being reported, for example, 'The experience of chronic pain in injured workers'. In general, titles should be concise and informative, enabling a prospective reader to identify the nature of the investigation. Immediately below the title will usually be the name(s) of the investigator(s) and affiliation (i.e. place of work).

The *abstract* is a short (approximately 250 words) description of the entire report. The purpose of this section is to provide the reader with a general overview of the communication. It should provide enough detail to enable the reader to decide whether the article is of interest. Writing an abstract is a challenging task because it should be precise and informative but at the same time very brief. When writing an abstract you should include:

1. A statement about previous findings that led to this research project.
2. The hypothesis and/or aims.
3. Methods, including participants, the research design and procedure.
4. A short description of what was found and how the data were analysed.
5. What was concluded and the nature of further research need.

Introduction

The introduction describes the planning stages of research, as discussed previously. A good introduction sets the stage for the research question by discussing the theoretical background of the problem under consideration and describing and evaluating the relevant research previously completed. The introduction thus serves as a link between past work and the present knowledge.

Generally, all aspects of the previous research literature cannot be covered in a relatively brief introduction required in a research paper; therefore the review of past research is done with a bias towards only those aspects of the problem that are of direct relevance to the report. In this way the research question can be derived in a logical manner. For this reason, the introduction generally starts by making a few general statements about the field of research, then focusing on specific research questions, aims and/or hypotheses.

Method

The purpose of the methods section is to inform the reader of exactly how the study was carried out. It is important to remember that the methods section should contain enough detail to enable another researcher to replicate the study. (Of course, replications may not be feasible for a unique event, such as a case study of a specific individual.) Conventionally, three subsections are reported: participants (or subjects), apparatus/tools and procedure.

Research Participants

Three questions must be answered concerning the participants: (1) who were they, (2) how many were there and (3) how were they selected? Specific information must be given concerning the participants, as results may vary from one sample to another. Ethical considerations also should be discussed here. Medical researchers generally refer to people in the sample as the ‘subjects’ and social scientists and qualitative researchers as ‘participants’.

Apparatus/Tools

A description of all equipment, questionnaires and standardized tests used in the research must be provided. If it is commercially available, the reader should be provided with the manufacturer’s name and the commercial identification of the equipment. Alternatively, if the equipment was privately made, the reader should be given enough information to allow replication. Measurements and perhaps a diagram will be necessary. In

qualitative research, equipment often includes devices used to record interviews.

Procedure

Once again, this section should provide enough information for other researchers to replicate the study. Details of how the research was carried out should include how research participants were assigned to groups, how many research participants were in each group, the experimental procedure used and a description of how data were collected.

Results

The results section presents the findings of the investigation and draws attention to points of interest. Raw data and statistical calculations are not presented in this section. Rather, the principles of descriptive and inferential statistics should be used to present the summarized and analysed data: graphs, tables and the outcomes of statistical tests are presented in this section. It is essential that all the findings are presented and that the graphs and tables are correctly identified. Qualitative researchers will present original quotes from participants and describe how analysis was conducted.

Discussion

The discussion section restates the aim(s) of the research study and discusses the results with reference to the aims or research hypothesis stated in the introduction. Did the researchers obtain the results expected? How do the present results relate to previous results and theories?

It is important to remember that the results of one study in isolation cannot make or break a theory or establish the efficacy of a novel practice. Thus the discussion should connect the findings with similar studies and especially with the theory underlying such studies. If unexpected results were obtained, possible reasons for the outcome (such as faulty design and controls) should be discussed. By doing this, the discussion will point the way to further problems that remain to be solved. Unconstructive, negative or unimportant criticism should be avoided so that the report does not end

with an unreasonably long discussion of possible reasons for the outcome. A brief, concise discussion is more appropriate.

In the conclusion, which is usually the last paragraph of the discussion section, the main findings are summarized and suggestions made for further research. For example, you may have demonstrated certain phenomena that may have implications for explaining broader concepts which can be empirically tested. You are therefore taking your findings and generalizing them to phenomena not directly tested in the present research.

References and Appendices

It is expected that all the literature discussed in the paper be listed in the references section. This enables the reader to evaluate the sources. Information on how references should be listed can be found in appropriate style manuals. Sufficient information must be provided for an interested reader to be able to identify and retrieve the sources. Additionally, a report may include labelled appendices. These might include a full description of questionnaires or other measuring instruments, raw data or statistical calculations if required.

The Publication Process

The formal knowledge representing the empirical and professional basis for your professional practice is in large part stored in journals, books and conference reports. Journals are published by appropriate professional associations, government departments or private publishing corporations. Having completed a research project, how does one publish it in a professional journal? After all, the value of research is negligible if it is not made public.

In general, the prospective author will:

1. Select a professional or scientific journal appropriate for the material.
2. Present the research report in a format required by the journal.
3. Send the completed manuscript to the journal editor.

The editor is generally a person of high standing in a given scientific or professional area. If the article is judged as being appropriate for the journal, the editor will send the article to two or more peer reviewers and, on the basis of the reviewers' reports, publish or reject the manuscript. Sometimes the reviewers recommend certain additions or changes which have to be made by the author before the manuscript is judged to be publishable.

Therefore when you read research publications in peer reviewed journals, you can be confident that experts have scrutinized the articles. However, discussed in [Chapter 22](#), the fact that a research study has been published in a reputable journal does not necessarily guarantee the truth of either the evidence or the conclusions.

Research and Evidence-Based Health Care

Research methods cover a wide variety of skills and techniques aimed at the methodologically valid investigation of questions of interest to the researcher. These methods of enquiry are not restricted to research laboratories, nor need they involve expensive equipment or large research teams. Rather, these methods imply an approach to stating and answering questions in any setting.

Research methods interact with health practices in multiple and mutually productive ways. An important general aim of this textbook is to discuss the relationships between theories, practices and the ways in which research methods contribute to improvements in health care. The term '*evidence-based medicine*' has been defined by Sackett et al as '... the integration of best research evidence with clinical expertise and patient values'. ⁵ Evidence-based health care is a contemporary movement aimed at ensuring that health services are based on the best available scientific and clinical evidence. The fundamental question that is explored in this book is 'what constitutes best research evidence to inform clinical practice?' The evidence is, of course, produced by research conducted in accordance with the principles of the scientific method. Clearly, if you expect to participate in the delivery of evidence-based health care you will need to understand research methods sufficiently for making informed and critical judgements

concerning the quality of the available evidence to drive high-quality and effective clinical practice. As outlined in [Chapter 23](#), systematic reviews of available evidence are very helpful to clinicians in keeping up-to-date in their clinical practices.

Finally, we need to keep in mind that the research is produced by the hard work of men and women who were undergraduate students, just like you. Ultimately, it will be up to you (as nurses, podiatrists, physiotherapists, speech pathologists and occupational therapists) to carry out the research that underpins the effectiveness, advancement and prestige of your profession. If not you, then who?

Summary

In this chapter we discussed the ways in which the scientific method is directly applicable to conducting research in the health sciences. The sequential steps of the research process follow a general problem-solving algorithm. Identifying these steps enables you to understand the rationale underlying the planning, conduct and interpretation of research. We outlined how the general format of a published research paper is associated with these steps. We also examined the style and structure of a research publication. The style involves clarity, accuracy and sufficient completeness for colleagues to understand or replicate the research project. Research is published in journals, which are generally edited by persons of high standing in the field. Every effort is made by editors to ensure the validity of the research published in their journals. The individual researcher also is bound ethically to report findings in an unbiased and truthful way. Although the format and style outlined in this chapter might seem rather arduous, poor presentation may destroy the intrinsic value of a research project.

References

1. World Health Organization. *Regional Guidelines for Continuing Medical Education (CME)/Continuing Professional Development (CPD) Activities* . New

Delhi: World Health Organization Regional Office for South East Asia; 2010.

2. Epstein L, Ogden J. A qualitative study of GPs' views of treating obesity. *Br J Gen Pract* . 2005;55(519):750–754.
3. Beaglehole R, Epping-Jordan J, Patel V, et al. Improving the prevention and management of chronic disease in low-income and middle-income countries: a priority for primary health care. *Lancet* . 2008;372(9642):940–949. doi: [10.1016/s0140-6736\(08\)61404-x](https://doi.org/10.1016/s0140-6736(08)61404-x).
4. Ng M, Fleming T, Robinson M, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* . 2014;384(9945):766–781.
5. Sackett D.L, Strauss S.E, Richardson W.S, Rosenberg W, Haynes R.B. *Evidence-Based Medicine: How to Practice and Teach EBM* . 2nd ed. Edinburgh: Churchill Livingstone; 2000.

SECTION 2

Research Planning

OUTLINE

Introduction

4. The Formulation of Research Questions

5. Sampling Methods and External Validity

6. Ethics

Introduction

The first stage of research involves developing the detailed plan for conducting the project. The project plan is documented in the *research protocol*. Before the research project is initiated, usually one or more research ethics committees (Medical Research Council 2018; National Health and Medical Research Council 2018) examine the protocol and funding bodies to ensure that the project plan conforms with general methodological and ethical principles. The three chapters in Section 2 aim to outline the basic considerations for the successful preparation of a research protocol. This section will help you build a picture of the overall research process.

The primary reason for proposing a research project is to obtain evidence that will, we hope, advance theory and practice in the health sciences and the efficacious patient-centred treatment of people with health problems (see [Ch. 4](#)). Before all else, we must be sure that we are asking the right research questions; that is, raising issues and problems which are crucial to progress to the development of high quality and effective health care. We must convince the experts who evaluate our proposal that the questions we are asking are of central importance. Asking the right research questions depends on being creative, for example, identifying unanticipated patterns in data, or the construction of novel theories that predict new, as yet unobserved facts regarding health care.

To justify the research proposal it is essential to write a literature review. The literature review is a summary and critical evaluation of previous research and theory relevant to the problem under investigation. The literature review provides both a conceptual background for our proposal and justifies the need for further evidence by identifying ‘gaps’ in our knowledge essential to solving the research problem. In contemporary health research it is now standard practice to use a systematic review

approach. The reporting of reviews is now a highly structured and systematic process. Systems such as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; <http://www.prisma-statement.org/>) are used to report what is known about particular research issues and questions. In a contemporary health research paper, if the paper does not itself use a systematic review, it is typical to report such reviews in the paper rationale and introduction. Systematic reviews may be found in many academic journals but there are also databases that are specifically focussed on systematic reviews. The most famous of these is the Cochrane Database of Systematic Reviews (<http://www.cochranelibrary.com/>). This database contains many thousands of systematic reviews about an exceptionally broad span of topics. The Cochrane library is actually mandatory reading for clinicians because it is a repository of the best evidence concerning clinical interventions. It is notable that the Cochrane Database includes a wide range of qualitative, quantitative and mixed reviews. This reflects the development of health research in including patient and clinician perspectives as valuable contributions to our knowledge base.

The way in which research study samples are selected is discussed in [Chapter 5](#). Selection of an appropriate study sample is crucial for the generalizability (or external validity) of your findings. The aim of sampling strategies is to select a representative sample of the study population. Also, we need to select a sample size which is sufficiently large to enable accurate inferences on the bases of sample data.

A project should not proceed unless it is judged to be ethically sound by an appropriate ethics committee (see [Ch. 6](#)). A research proposal is judged to be ethical if it conforms to the rules of moral conduct and values of the community and the evaluation criteria used by the committee to consider this issue. These rules and values are made explicit in documents and research publications representing the standards of professional groups and of institutions (e.g. hospitals, universities and research councils). The research protocol must describe the research in sufficient detail the degree of burdens imposed on the participants.

The Formulation of Research Questions

Introduction,
Identifying Research Problems,
 Environmental and Social Changes,
 Demographic Changes,
 Scientific and Technological Advances,
 The Cost-Effectiveness of Health Services,
Literature Reviews and Research,
Formulating the Research Question,
Methodological Considerations,
 Deciding Between Using Qualitative and
 Quantitative Methods in Your Research,
 Research Participants and Sampling,
 Study Design,
Ethical Issues and Procedures,
Aims and Hypotheses,
Research Proposals,
Summary,

Introduction

The planning of a research study begins with the identification of a research problem, followed by the formulation of the research question(s). In the ‘Introduction’ section of a research paper you will find the research question guiding the investigation. The formulation of a research question is the outcome of a complex, iterative process. Researchers refine and focus the question as they plan the stages of the research project, as shown in [Fig. 4.1](#). The aims of this chapter are to:

1. Describe the process for identifying research problems.
2. Discuss the use of literature reviews for developing research questions.
3. Describe the relationship between formulating research questions and methodology.
4. Explain the relevance of ethics and resources to the statement of specific aims and hypotheses.
5. Describe what constitutes a research proposal.

Identifying Research Problems

Research in health and medical science often is focused on identifying effective interventions for preventing or treating health problems. Many research problems are derived from improving current practices, as discussed in Sackett et al. ¹ Economic considerations, that is, the availability of resources, also can have a strong influence on research planning. For example, you may have designed a qualitative research project that involves 100 ‘in-depth’ interviews with persons suffering from a disorder. Say that you are an ‘honors’ student and have only 1 year and a limited amount of funds to complete the project. You would be advised to reduce your sample size or, if this is not possible, change your topic. Ethics or funding bodies will only approve projects which are realistic – that is, those which can be completed feasibly with the available resources.

Thus research planning is a process through which we transform our initial ideas into well-planned, ethical and realistic research projects ([Table](#)

4.1). Experts judge the feasibility of the project before permission is given to proceed.

However, there are numerous other research problems, such as gaps in our knowledge of the causes and consequences of health problems and opportunity to improve the quality of our practices. Let us look at examples of some research problems.

Environmental and Social Changes

As the biological and social environment changes, new health problems may emerge. For example, abundant food supplies and increasingly sedentary lifestyles have led to a high prevalence of obesity in many communities. The Global Burden of Disease studies (<http://www.healthdata.org/gbd>) conducted by the international Global Burden of Disease consortium is an exceptionally comprehensive epidemiologic study global program in which obesity and other disease-burden risk factors have been studied. Many publications in the *Lancet* and other journals report this work. We consider a good knowledge of this work as a valuable tool in the clinician's clinical toolkit. It is well worth a comprehensive read.

Literature review:
identify current knowledge
and gaps in knowledge



Research problems
and questions



Aims and hypotheses



Methods:
• Sampling
• Design
• Data collection



Ethics and available resources

FIG. 4.1 Reiterative process for formulating research questions.

Obesity is a serious risk factor for numerous health problems, including type 2 diabetes and coronary heart disease. The research problem of interest to a researcher may be the lack of effective interventions for preventing and reducing obesity in the community.

Another example is the health effect of climate change, a process which is predicted to affect population health. The environmental changes (in particular the prediction of global warming) have led to identifying new research problems and questions about the effect of climate change. Many countries have research priorities based on the prevalence of diseases as identified by epidemiologic studies (see [Ch. 8](#)). Public and political concern about the ‘burden of disease’ for a community results in funding bodies preferentially channelling resources to researchers in these areas.

Demographic Changes

Changes in population characteristics, such as the increasing number of elderly people, have shifted the health needs of communities in many countries. There is increased research activity concerning the health problems associated with ageing and the ways of designing and delivering effective and efficient health services for elderly patients. Also, some countries have an increasing number of immigrants and asylum-seekers who have health needs different than the rest of the population. Relevant research questions have been asked regarding language or sociocultural barriers which prevent migrants from accessing appropriate health services. In general, research problems arise from the limitations of the current health services being offered to a population with changing needs.

TABLE 4.1

Categories of practice	Typical research questions
Improving diagnosis and screening	Does this new diagnostic test accurately discriminate between people who suffer from Alzheimer's disease and multistroke dementia? Does this new genetic test accurately identify people who are likely to suffer from epilepsy?
Producing and evaluating strategies for predicting health outcomes	Does the immunization of children result in increased incidence of autism? What is the fatality rate among people who were exposed to high levels of mercury?
Devising and evaluating treatments	What are the comparative benefits of hydrotherapy and nicotine replacement therapy for smoking cessation? Does aspirin reduce the incidence of stroke?
Harm minimization	What are the comparative infection rates after surgery across a group of metropolitan hospitals? Do hypertension medications increase the risk of kidney failure?

Scientific and Technological Advances

There are continuing and remarkable discoveries in the biological, behavioural, social and information sciences that can be applied to improving the assessment, treatment and prevention of illnesses. Advances in electronics and information technology have created opportunities in a variety of health professions, devising new diagnostic techniques or data collection and management strategies. The research problems arise from the need to *translate* pure research into practical solutions. An example of this is stem cell research, which has led to spectacular discoveries in the biological sciences. Translational research in this field of regenerative medicine aims to develop cellular therapies for the treatment of a variety of disorders such as heart disease, diabetes and Parkinson's disease. [Chapter 24](#) examines the rationale for translational research.

The Cost-Effectiveness of Health Services

Economic pressures from increasing health costs require that service providers are accountable for the way in which they use available resources. Program evaluations for services are now routinely incorporated into service delivery. Evidence from this research is necessary to demonstrate that the resources are being used in the best way and that the services are meeting the needs of patients and the community. Although often a source of stress for health professionals and administrators who participate in them, well-designed evaluation programs can contribute to the development of cost-effective, 'value-for-money' services to the community. For

example, if you can demonstrate that program A and program B provide the same benefits but A costs half as much as B to deliver, then you will have made a significant contribution to health care. The money saved can be used to provide better services to other patients with other problems.

We hope this discussion demonstrates that there are innumerable research problems waiting for you when you graduate! Perhaps you will take up the challenge and by asking the right questions will produce research which will contribute to the advancement of health care for the community.

Literature Reviews and Research

Before embarking on the design and conduct of a research project, the investigator must review previous work and publications relevant to the aims of the intended project. (The process of literature review is described in detail in [Ch. 23](#)). This process is essential, both for providing the appropriate background and context for the investigation and for justifying the investigation in contributing to existing knowledge. It is potentially a waste of money to inadvertently duplicate very similar research (although deliberate replication of previous studies may be a legitimate activity if there are uncertainties about the validity of a previous research). Another reason for reviewing the literature is to examine how previous researchers have approached sampling, design, data collection and analysis decisions in their research. There are benefits in reviewing how previous researchers succeeded or failed in achieving their research and clinical objectives.

There are several standard approaches to identifying research papers of interest to a specific research study or problem. An example is PICO (see <http://pubmedhh.nlm.nih.gov/>), an acronym for:

- P** = Target Population/Patients
- I** = Intervention (or type of exposure)
- C** = Comparison (alternative intervention)
- O** = The Outcome measures

By examining the keywords, title and abstract of a paper, you can see how these criteria apply. Imagine that you are interested in identifying papers that investigate the efficacy of neural transplantation for treating

Parkinson's disease. PICO has been used for a long time. For example, you should be able to find the paper, 'A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease' by Olanow et al. ² Here:

P represents patients with Parkinson's disease.

I refers to the bilateral foetal nigral transplants.

C refers to the placebo surgery group.

O refers to outcomes on standardized tests.

Literature searching may be carried out at appropriate research centres or libraries where scientific and professional journals are stored and, of course, by using the Internet. In health research, as opposed to historical research, almost all searches use Internet search engines and electronic journals databases. There are powerful web-based search methods that can simplify the search. The most widely used search tool for health research publications is probably PubMed, the web-based Public Medline (see <http://www.ncbi.nlm.nih.gov/entrez/>). Professional library staff can help you locate the relevant literature. However, the critical evaluation of the literature depends on the application of research methods for the identification of controversies or 'gaps' in the available evidence (see [Ch. 23](#)). A detailed and critical review of the literature enables us to ask the questions most relevant to planning the kind of research that will enable us to address the research problem.

Formulating the Research Question

To formulate a research question, we need to investigate what is and what is not known in the area (i.e. to identify what we do and do not know when we understand the existing published literature). By stating the research problem and reviewing the literature, we are able to devise a research question. The research question will guide the investigation and guide us in producing the evidence required to solve the research problem. Inspiration for research questions may stem from direct personal experience of working with patients and literature review and analysis. The critical evaluations to an issue or problems may raise questions suitable for a research study. The identification of research questions is a creative act. It draws upon

consideration of previous relevant work in the area under study. This work may be reviewed and the review may give rise to questions that have not been answered by previous work.

Methodological Considerations

The precisely stated research question requires an examination of the proposed methodology of the project. To illustrate the methodological issues, let us look at an example based on a publication titled ‘Ribavirin for Crimean-Congo hemorrhagic fever: systematic review and meta-analysis’.³ The health problem is the Crimean-Congo haemorrhagic fever (CCHF), which is a potentially lethal viral disease. There is a drug called ribavirin which has been approved by the World Health Organization (WHO) and has been used with apparent success to reduce the mortality rate associated with CCHF. A review by Soares-Weiser et al,³ identifies a research problem and a possible research question regarding the efficacy of ribavirin:

‘No clear message of benefit is available from the current data on ribavirin as observational data are heavily confounded, and the one trial carried out has limited power. However ribavirin could potentially have benefits in this condition...’

Let us imagine that we are scientists working in public health and that we want to contribute to solving this problem. The first thing to do is to formulate a precise research question and, if appropriate, a research hypothesis derived from the research question. To do this, we need to plan the data collection process for conducting the investigation.

Deciding between Using Qualitative and Quantitative Methods in Your Research

An essential consideration in framing the research question is whether you will use a qualitative or a quantitative research method. Quantitative research is normally structured so as to test a research hypothesis. Hypotheses are propositions about relationships between variables or

differences between groups that are to be tested. Qualitative research does not usually test hypotheses. It is concerned with understanding the meaning of how people experience specific issues. However, most qualitative researchers work with clear aims and questions to guide their investigations. Some research questions require both qualitative and quantitative methods.

For our hypothetical example, we would select a quantitative approach. The reason is that we intend to test the *causal* hypothesis that ribavirin, an antiviral drug, is effective for treating CCHF. If we were investigating the experiences of patients taking this treatment, we would have selected a qualitative approach. However, we are looking at the efficacy of a treatment. The next step will be to refine our research hypothesis, taking into account the sample and proposed design.

Research Participants and Sampling

To focus our hypothesis, we need to define the population we wish to study. In the ribavirin example, the study population consists of people living in regions where CCHF is a public health concern. According to Soares-Weiser et al,³ recent outbreaks have occurred in Turkey and Iran. Because of both clinical and pragmatic reasons (see the next section), you might select people living in Turkey as your population of interest. We will look at details of sampling in [Chapter 5](#).

Study Design

In [Section III](#) we will examine research designs in detail. As discussed previously, we have selected a quantitative methodological approach for our hypothetical investigation. There are several types of research designs which fall under quantitative methods; the most basic distinction is whether they are *experimental* or *nonexperimental observational* designs (see [Section III](#)). In our example, we are investigating a causal hypothesis; therefore we would select an experimental design. As discussed in [Chapter 7](#), a randomized controlled trial (RCT) would be the preferred design for establishing the efficacy of ribavirin.

In an RCT, the participants are assigned (by chance or random method) to a treatment (active intervention) or control (non-intervention) group, and

then given the intervention or, in clinical contexts, the treatment. In our example, the treatment is the quantity and administration of the ribavirin. We also need to make a decision about the type of control group we will use (i.e. no treatment, placebo treatment or current standard treatment other than ribavirin). The intervention and control groups are then observed and compared. At this point, we must consider how we are going to measure outcomes for enabling us to make a comparison. The selection of outcome measures is discussed in detail in [Chapters 11 to 14](#). In this case, since CCHF is a potentially fatal condition, we should use mortality rates as the indicator of treatment efficacy.

Ethical Issues and Procedures

When conducting health care research, a primary concern is the well-being and dignity of the research participants (see [Ch. 6](#)). Ethical considerations will influence the sample selected, the design of a project and, ultimately, the research question which will be posed. Additionally, the available resources such as time and money will determine the scope of the research project. In our present hypothetical example, we may have decided to:

- Select the sample only from Turkey for budgetary reasons.
- Use a best-available treatment for the participants in the control group for ethical reasons.
- Use mortality as an outcome measure for treatment efficacy.

All studies need to be ethically cleared by a properly constituted human research ethics committee (see [Ch. 6](#) for a detailed discussion of ethics processes).

Aims and Hypotheses

Having considered the methodology and ethics, we are able to refine our research question as follows: ‘Is ribavirin more effective than the best-available current treatment for reducing mortality in people with CCHF living in Turkey?’ Our aim is to collect evidence showing the efficacy of

CCHF. In general, the aim of a research project is to produce evidence for answering the research question.

Because we have selected a quantitative method to investigate the question, we can also state a research hypothesis: ‘Ribavirin treatment will reduce mortality more than best-available treatment in people with CCHF living in Turkey’. The obtained results will enable us to make a decision as to whether the research hypothesis is supported by the data we collect. This decision will enable us to answer the research question and contribute to solving the original research problem. You will recall that the research problem was an uncertainty regarding the efficacy of ribavirin. [Fig. 4.1](#) represents the steps that are followed to develop detailed and precise aims and hypotheses.

Research Proposals

By the time we have stated our precise aims or hypotheses, we should have decided the way in which the data are to be collected. This information is written up as a research proposal. A research proposal is an explicit statement of how and why you intend to conduct a research project. A research proposal will contain essential information:

- Identification of the research problem
- Literature review
- Research aims and hypotheses, which must take into consideration the:
 1. participants to be selected,
 2. data to be collected,
 3. research design/s,
 4. ethical considerations and
 5. estimated cost.

The research proposal is circulated to various individuals and committees in order to obtain critical feedback from supervisors, institutions and your fellow students concerning the practicality and quality of your research project. Most importantly, the research proposal is the basis for writing a detailed report to relevant ethics committees (see [Ch. 6](#)).

It often is desirable to carry out a small-scale preliminary study called a *pilot study*. This is an economical way of identifying and eliminating potential problems before you do the large-scale (and expensive) full study. Case studies and so-called safety and efficacy studies often are employed before RCTs are undertaken (see [Ch. 7](#)).

Summary

The planning of a research project requires the transformation of preliminary thoughts on solving a research problem into clearly stated research questions and aims. To achieve this, the researcher should review the relevant literature and evaluate ethical considerations and economic constraints in conducting the investigation. Next, an appropriate research strategy has to be formulated. Having considered the stages of the research project, the researcher is in a position to state precisely the aims or hypotheses being investigated. Before data collection begins, an ethics committee must scrutinize the research proposal to resolve methodological problems and to ensure the implementation of ethical research.

References

1. Sackett D.L, Strauss S.E, Richardson W.S, Rosenberg W, Haynes R.B. *Evidence-Based Medicine: How to Practice and Teach EBM* . 2nd ed. Edinburgh: Churchill Livingstone; 2000.
2. Olanow C.W, Goetz C.G, Kordower J.H, et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Ann Neurol* . 2003;54(3):403–414.
3. Soares-Weiser K, Thomas S, Thomson G, Garner P. Ribavirin for Crimean-Congo hemorrhagic fever: systematic review and meta-analysis. *BMC Infect Dis* . 2010;10:207.

Sampling Methods and External Validity

Introduction,
Basic Issues in Sampling,
Representative Samples,
Sampling Methods,
Sample Size,
Sampling Error,
Sampling in Qualitative Research,
Purposive Sampling,
External Validity and Sampling,
Ecological Validity,
Summary,

Introduction

Research in the health sciences usually involves the collection of information from a sample of participants, rather than the entire population in which the investigator is interested. Studies that involve an entire population or group are called census studies, but these are relatively rare

and generally prohibitively expensive to perform. Thus a sample is drawn from the target population and studied because it is usually impossible or too costly to study entire populations. For instance, when individuals who have conditions such as diabetes, cerebral palsy or emphysema are being studied, it is not possible to study everyone with the disease because of the large size of such populations. Also because many people do not seek treatment or may have been incorrectly diagnosed, it may not be possible to identify all members of the entire population. Therefore in most research, the researcher studies a subset or sample of the target population and then attempts to generalize the findings to the population from which the participants were drawn. This general principle applies to both qualitative and quantitative studies.

The aim of this chapter is to examine ways in which samples can be drawn to permit the investigator to make valid generalizations from the study sample to the target population.

We will also consider the question of generalizing the findings of an investigation to other samples and situations. This is referred to as the *external validity* or *generalizability* of a sample.

The specific aims of this chapter are to:

1. Define what is meant by sampling and representative samples.
2. Outline the relative advantages and disadvantages of commonly used sampling strategies for both qualitative and quantitative methods.
3. Discuss the relationship between sampling error and sample size.
4. Examine the concept of external validity for generalizing research findings.

Basic Issues in Sampling

A *population* is defined as the entire set of persons, objects or events that the researcher intends to study. That is, the population is defined in terms of the specific research question being investigated. We must examine very closely the research questions being asked to clarify the population which is being studied in a research project. In other words, the population can be any group of people or set of objects. For example, consider the statement

from a textbook for health professionals working with people with Parkinson's disease (PD): ¹

*Levodopa is the most effective medication for PD and almost all patients will need to be on levodopa during the course of the disease.*¹
(p26)

This statement refers to the actual and predicted clinical responses of millions of people with PD to a range of drugs which increase acetylcholine activity in the brain. It must be evident to you that the proposition just given is based on evidence from a minute proportion of patients with PD. You might recall that the logic for generalizing from a limited set of observations is called *induction*. Induction was defined as the logic of producing general statements on the basis of specific observations. In the context of research, the 'limited set of observations' refers to sample data collected in the course of the study. The process of generalization entails summarizing sample data and then drawing inferences from sample data about the probable true state of the population. We will also examine this concept in the statistical chapters in Section V and VI.

A *sample* is defined as a subset of the population. When we talk about a population we are not only referring to the individuals who constitute the population but also the factors or variables which we observe or measure in these people. For the example given here, the sample is not simply the people with PD but also the changes in their symptoms. The data for the symptomatic changes in people under the influence of drugs which cause increased acetylcholine constitutes the actual sample.

In the context of scientific research, we call the process of generalizing from sample data to the population *inference*. Inference is probabilistic; that is, we are not absolutely certain that what we find in our sample is true for the population. To improve the probable truth of our inferences we must ensure that our samples are representative of the population.

Representative Samples

There are a variety of different ways by which one can select the sample from the population. These are called *sampling methods*.

The ultimate aim of all sampling methods is to draw a representative sample from the population. The advantage of a representative sample is clear: one can confidently generalize from a representative sample to the rest of the population without having to study the entire population. Jones and Attia (2017) in their aptly named paper ‘Sampling: how you choose people is as important as how you analyse their data’ note that ‘if the research design is poor, no amount of complex statistical analysis can extract useful information from the data collected’. ² (p67)

A representative sample is an accurate but reduced representation of the population being studied (Table 5.1). Ideally, a representative sample contains in a miniature form all the relevant factors or variables in the population in the same proportions and magnitudes as in the original population. If a population is not representative then it is *biased*. Biased samples generate *sampling errors*, which are defined as the magnitude of the difference between the true state of the population and the state of the sample. So if the average age of the targeted population was 45 years but the study sample had an average age of 50 years, then the sampling error for age would be 5 years.

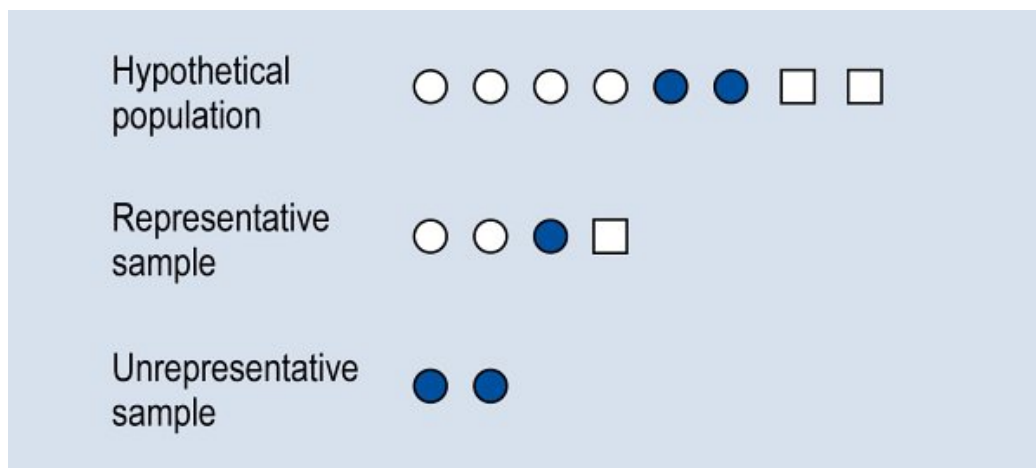


FIG. 5.1 A simple representative sample.

Fig. 5.1 illustrates a hypothetical population composed of three different shapes.

A representative sample is a precise miniaturized representation of the population. An unrepresentative, or biased, sample does not adequately

represent the key groups or characteristics in the population, and this may lead to erroneous conclusions about the state of the population. The first sample shown in [Fig. 3.1](#) is representative because it has the same proportion of the three shapes as the population. The second sample is obviously biased and far too small even to correctly represent the population.

Sampling Methods

The selection of the appropriate sampling method depends on the aims and resources of the researchers. For instance, if someone is designing a very expensive health or social welfare program on the basis of a survey of client needs, it is imperative that a good sampling method is used and a representative sample of the clients be obtained so that appropriate conclusions may be reached about the population. Good sampling methods tend to be more expensive and more difficult to implement than poor methods, but, arguably, they are worth it because of the effects upon the soundness and quality of the research.

Incidental Sampling

Incidental sampling is the cheapest and easiest sampling method for conducting clinical studies. It involves the selection of the most accessible and available members of the target population. For example, a researcher who stands in a street in the business centre of London (the City) and quizzes people about their health status is practicing incidental sampling. However, it is quite likely that this sample would not be representative of the target population. There probably would be an overrepresentation of businessmen and white-collar workers with higher salaries, and an underrepresentation of farm and factory workers. The sample is likely to be unrepresentative and biased as a sample, if the target population was defined as adults living in the United Kingdom.

Table 5.1**Examples of Populations and Samples**

Population	Possible sample selection
All working podiatrists in a state or province	50 podiatrists selected for a study of job satisfaction
All working pathologists in a state or province	25 pathologists selected for detailed tax evaluation by an inspector
The temperature of a patient during a 24-h period	Hourly measurements of a patient's temperature recorded by staff
Stuttering in a child's speech	Number of stutters made during 5 min of reading a standard piece of material
All patients in a state with frontal lobe damage	30 patients with frontal lobe damage selected for evaluating a rehabilitation program
All surgical gauzes held by a given hospital	10 gauzes selected by a bacteriologist to test for sterility

A further example of incidental sampling might involve a researcher surveying the needs of a group of children with spina bifida at a local community health centre. Their measured needs may be representative of those of other children with spina bifida but then again they may not if these children are not typical of the wider population of children with the condition.

Thus incidental sampling is cheap and easy to implement but may give a biased, sample target population.

Quota Sampling

Sometimes it is known in advance that there are important subgroups within the population that need to be included in the sample. Two such important groups within the human population are males and females. Furthermore, it

is known that they occur in the ratio of approximately 49:51 in the general population. The researcher might decide that it is very important that the sexes are proportionally represented in the sample. Thus the researcher would set two quotas of 49 male and 51 female respondents in a sample of 100. This is still a form of incidental sampling but has some significant advantages over simple incidental sampling because the study sample's composition on this key demographic variable is guaranteed to match that of the target population.

Table 5.2

	Blue collar	White collar	Not employed 'no collar'	Total
Male	19	21	9	49
Female	15	15	21	51
Total	34	36	30	100

More sophisticated examples involving more than two groups can be accommodated as shown in [Table 5.2](#). We can see from [Table 5.2](#) that if the sample were to be representative regarding both sex and occupational status, in a sample of 100 people, we would need 19 blue-collar male workers, 15 blue-collar female workers, and so on.

Quota sampling still has a number of shortcomings: before it can be used, one has to decide which population groups are likely to be important to a particular question and have access to data for the exact proportions of the various groups. Sometimes we may not know these numbers. Also, the members of the sample within the quotas still are incidentally chosen. The blue-collar males, for example, selected in a city centre on a weekday may still be quite different from those working elsewhere in regional, remote or even suburban areas. However, quota sampling is better than simple incidental sampling in that sampling error is reduced by the sample being accurate on at least a few key variables.

Random Sampling

Random sampling is one of the best but probably more expensive methods to implement in drawing a study sample. A random sample is one in which all members of the population have an equal chance of selection. Thus a

random sample is more likely than an incidental sample to be representative of the relevant population.

The procedure for drawing a random sample involves:

- Constructing of a list of all members of the population.
- Using a method such as dice, coins, a hat or random number tables to select randomly from the list the number of members required for the sample.

A simple example of a random sample is provided by a common raffle, where names on (preferably!) equal size papers are put in a hat, shaken and selected 'blind'. Many national and state lotteries use numbered balls that are drawn randomly from a barrel. Another way to draw a random sample is to construct a list of all the members of the population and assign a number to each element. Then a table of random numbers, generated by a computer, can be used to select a random sample.

Cases that are selected from a list using a random selection method constitute a random sample. Sometimes, because of issues such as refusal to participate in the study, dropping out and failure to satisfy the sample inclusion criteria, it is necessary to select replacements for some cases that are unavailable for the study. This can introduce bias into the sample, as the people who refuse may differ consistently from those who accept or volunteer. This effect on sampling validity is sometimes called the volunteer effect.

However, random sampling methods have a number of important advantages over incidental or methods that are not random:

- Because the exact sizes of the sample and population are known, it is possible to estimate exactly how representative the sample is: that is, the size of the sampling error. This cannot usually be done with sampling methods that are not random.
- Because random samples are usually more representative than samples that are not random, the sample size needed for good representation of the population is smaller.

The major disadvantages of random sampling methods are:

- The researcher potentially needs to be able to list every member of the population. Often this is impossible because the full extent of the population is not known. For example, it would be very difficult to sample randomly from the population of Canadians with coronary heart disease because no such list exists. Even if it did exist, it would be constantly changing.
- Cost. It is usually easier to use conveniently available groups. Random sampling usually involves considerable planning and expense, especially with large populations.

If the study involves a telephone-administered survey, one might use a Random Digit Dialling method. However, there is certainly good evidence now that use of such methodologies will deliver biased samples that are not truly 'random' and unbiased.³ This has been found to be true in low- and middle-income countries with different patterns of access to telephones to high-income countries.⁴ Also, significant differences have been shown in the results obtained from mobile versus landline phone services, to the extent that many researchers now use quota sampling to establish quotas for the different types of telephone.

Stratified Random Sampling

Stratified random sampling involves the same approach as quota sampling with set quotas from specific sub-groups, except that each quota is filled by randomly sampling from each subgroup, rather than by sampling incidentally. For example, if one was drawing a sample stratified with respect to sex, one would prepare a list of all females and all males in the target population and then sample randomly from these lists with the numbers of each group in the sample corresponding to the population proportions. These groups collectively are called the *strata*.

The advantages of stratified random sampling are:

- All the important groups are proportionally represented.
- The exact representativeness of the sample is known, which has important statistical ramifications.

The disadvantages are:

- A list of all members of the population, their characteristics and the proportions of the important groups within the population need to be known.
- Cost.
- The gain in sample accuracy is usually small compared with simple random sampling.

Area Sampling

In area sampling, one samples on the basis of location of cases. For example, on the basis of census data, the investigator may select several areas in a city or county with known characteristics, such as high or low unemployment rates. The areas could then be further divided into specific streets and the occupants of, say, every third house contacted for participation in the study. In other words, the locations are randomly selected and then one interviews the occupants of those locations. This can be a very effective, inexpensive method of sampling in social surveys. It does not require a list of the individual members of the population, merely the location where they live. Indeed, one does not even need a list. However, in recent considerations of 'randomness' of sample selection, area samples are not considered truly random. Time of day and year can radically affect availability in neighbourhoods.

Systematic Sampling

Systematic sampling involves working through a list of the population and choosing, say, every 10th or 20th case for inclusion in the sample. It is not a truly random technique but it will usually give a representative sample. It is based on the (usually justified) supposition that cases are not added to the list in a systematic way which coincides with the sampling system. Provided a list of cases is available, systematic sampling is an easy and convenient sampling method. In clinical practice, we use systematic sampling when measuring temperature and blood pressure every hour.

Sample Size

One of the most poorly understood aspects of sampling is the number of cases that should be included in a study sample. The whole issue of sampling is conceptualized somewhat differently for qualitative and quantitative studies, although the issue of to what extent the results of a study are generalizable is the same regardless of the study design.

It is obvious that in one sense the more cases selected the better (or the better the sample representativeness), but the costs associated with the sampling and data collection must be weighed against the greater generalizability that is generally associated with larger samples. Also, some studies may involve discomfort, pain or even danger to patients or laboratory animals. Therefore in this case it is ethically and logistically desirable to ensure that no more than the bare minimum of participants is used to achieve the desired sample accuracy. Health researchers ‘walk a tightrope’ in deciding the optimum sample sizes for their studies. However, there are some principles available to guide the researcher.

First, let us say quite definitely that there is no magic number that we can point to as the ‘optimum’ sample size. Rather, the optimum sample size depends on the characteristics of an investigation, in the context of which the sample is drawn. [Fig. 3.1](#) shows that the sample size of 215 is too small to represent the hypothetical population. In general, the optimum sample size is one which is adequate for making correct generalization from the sample to the target population. Let us discuss these issues by introducing the concept of *sampling error*.

Sampling Error

In the quantitative research framework, sampling error is the discrepancy between the true population parameter and the sample statistic. ⁵ For example, if I happen to know from census data that the actual average age of males in a district is 35 years and the average age of a sample of males I have surveyed from the district is 30 years, then I have a sampling error of 5 years. However, if we do not know the actual population parameters (which is commonly the case), we can only estimate the probable sampling error.

Sampling error is related to sample size (n) by the following relationship:

$$\text{Sampling error} = \frac{1}{\sqrt{n}}$$

This equation indicates that the greater the sample size, the smaller the probable sampling error.

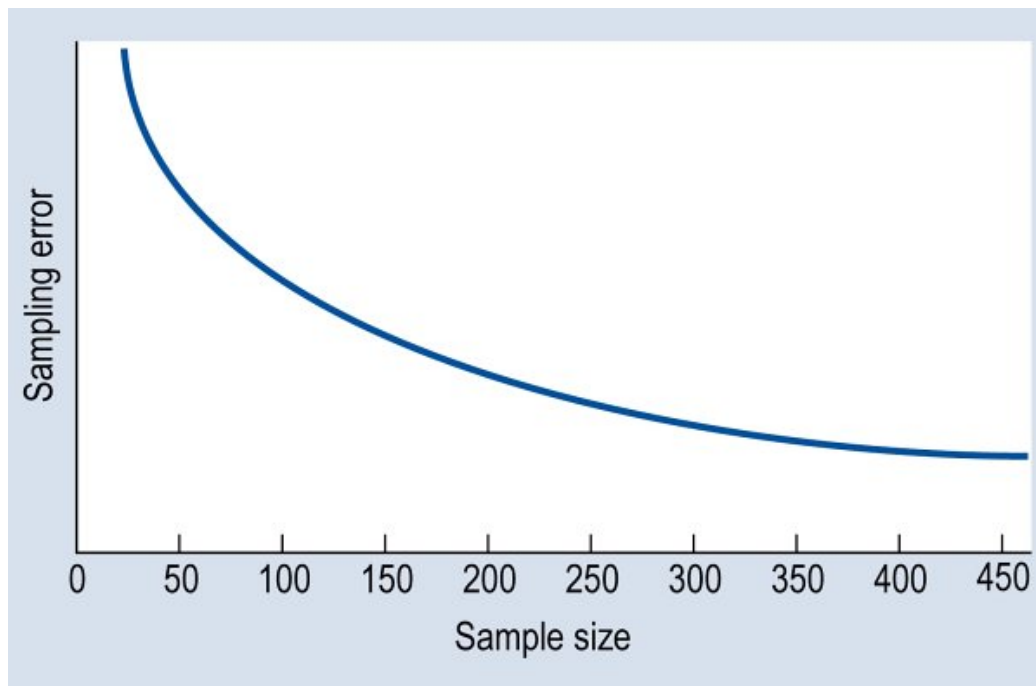


FIG. 5.2 The relationship between sample size and sampling error. The scaling but not the form of the curve will alter with the variability of the data.

From this relationship it can be seen that doubling the sample size would only result in a reduction of the error by a factor of the square root of 2 (1.414). Similarly, a ninefold increase in sample size would result in only a threefold reduction in the sampling error. [Fig. 5.2](#) illustrates this point by showing the graphical relationship between probable sampling error and sample size.

It can be seen from this graph that not much is gained from a sample size of more than, say, 300. Yet the cost of the sampling and data collection can be very high with large numbers for relatively little gain in reducing sampling error. In some research situations, even large probable sampling errors have relatively little potential influence on our decisions. In such situations, we can live with a relatively small sample size. In other situations, we need large samples to justify our confidence in the truth of our decisions.

As an illustration, suppose that we are attempting to predict the outcome of an election fought between two political parties: A and B. A representative sample of 100 respondents is polled before the election. Say the outcome is as follows:

Intends to vote for A: 25%.

Intends to vote for B: 75%.

Estimated sampling error: 10%.

In this instance the estimated sampling error is very small in relation to the size of the effect (i.e. the difference between the percentage of intended votes for the two parties). Assuming that the respondents were truthful and don't change their minds before the elections, we can predict confidently that political party B will romp into government. Increasing sample size, say to 10 000, would enormously increase the cost of the survey. The corresponding reduction in sampling error would not justify this cost, as we would still come to the same conclusion. However, say the following sample statistics were obtained in the preelection poll of 100 respondents:

Intends to vote for A: 48%.

Intends to vote for B: 52%.

Estimated sampling error: 10%.

Now, in relation to the apparent size of the effect, the same level of estimated sampling error is too large to make a decision concerning which party is likely to win the election. The poll would have to be repeated using a substantially increased sample size to reduce the sampling error.

This example illustrates the notion that the adequacy of the sample size is affected by the specific situation that we are trying to research. A sample size which is adequate in one situation may be inadequate in another. One benefit of a previous pilot study is that it allows the researcher to estimate the size of the phenomenon under study, and thereby make a more educated guess concerning the sample size required. The concept of statistical 'power' or the ability to detect real effects or differences is discussed later in this text.

Sampling in Qualitative Research

Qualitative research shares with quantitative research a concern with the extent to which findings from one study can be generalized to other settings and people. Therefore the representativeness of the information obtained from the sample is important in qualitative research. However, there are differences in the way in which a representative sample is conceptualized in qualitative research.

The key difference is the way in which the population is defined. In quantitative research we make the assumption that there is a true knowable state of the population. The true state of the population is represented by an actual parameter, such as '49% of the Australian population are females' or 'the average IQ of Canadian high school students is 104'. Qualitative research is not concerned with measuring quantities or counting frequencies with categories but with the experiences and meanings of these experiences for individuals living in culturally defined communities. Sampling in qualitative research is referred to as *purposive*; the researchers deliberately select the participants who are best placed to provide the information for understanding the personal meanings of health-related events. For example, what is it like to survive a heart attack? What does it mean for patients to undergo transplantation surgery? The sample provides the information which enables us to understand the emotional processes of coping with a heart attack or the psychological demands of being an organ recipient.

Let us consider an example of this approach. Say that you are intending to conduct qualitative research for clarifying issues relevant to the problem of family or domestic violence. The first step would be to clarify the specific aspect of the problem you were studying and define the issues in

which you were interested. For example, say that you wanted to collect data concerning the effects of physical violence on the lives of women. A key issue in defining the population is the cultural and historical setting in which it is enmeshed. The moral, legal and health-related problems involved in domestic violence vary across cultures and change with time. In Western societies it is only relatively recently that violence in domestic settings has been recognized as a serious crime, and this influences the experiences of the people involved in the events. Therefore it is essential to carefully define the cultural characteristics of the group being studied and to keep in mind that we must be very cautious in generalizing the findings from one study to other settings and people.

Purposive Sampling

Qualitative sampling is designed to be purposive; that is, to select cases for inclusion in the research that are likely to be illustrative of particular issues or circumstances. Patton ⁶ proposed a typology of sampling that is widely used in qualitative research. [Table 5.3](#) shows four of the commonly used purposive sampling methods. Interested readers can review Rice and Ezzy ⁷ for more detailed information.

To illustrate qualitative sampling, we selected four strategies. Considering the example of physical violence against women in domestic settings, what would be the researcher's purpose for selecting the following strategies?

1. *Extreme or deviant case sampling.* We would select respondents who had experienced severe, even life-threatening physical violence. Alternatively, we could select women who had experienced minimal violence (such as verbal abuse). Both extremes would illustrate the consequences of violence from the victim's perspective and the understanding of these issues could contribute to developing sensitive and effective interventions.
2. *Maximum variation sampling.* Here we would select across all the categories and severities of domestic violence which have been previously identified. The purpose here is to identify the common

themes emerging from the victims' experiences. This approach would also increase the understanding of how different types of violence might result in varying outcomes from the persons' perspective.

3. *Snowball or chain sampling*. The assumption here is that people sharing similar important experiences might become aware of each other in a community. In particular, key informants who are sympathetic to the researcher's work might be helpful in identifying 'information rich' cases. For example, people who are actively involved in victim support groups would be a good source for further contacts with possible participants. Additionally, such key informants can act as 'gatekeepers', protecting participants from unsafe researchers. However, 'snowball' sampling can lead to the biased selection of participants and an inaccurate representation of the issues.

Table 5.3

Commonly Used Purposive Sampling Methods

Sampling method	Brief description of sampling strategy
Extreme or deviant case sampling	The cases in these methods are chosen because of their deviance and the hope that this deviance may illustrate issues about more regular or mainstream cases
Maximum variation sampling	Purposefully picking a wide range of variation on dimensions of interest Documents unique or diverse variations that have emerged in adapting to different conditions Identifies important common patterns that cut across variations
Snowball or chain sampling	Identifies cases of interest from people who know people who know people who know what cases are information-rich (i.e. good examples for study, good interview participants)
Theory-based or operational construct sampling	Finding manifestations of theoretical construct of interest so as to elaborate and examine the construct

4. *Theory-based sampling.* A researcher might hold the theory that the occurrence and experience of family violence might be shaped by patriarchal traditions. Here the researcher might purposely select couples who have more-or-less equal relationships and couples where there has been an imbalance of power. The theoretical construct of interest might be ‘power’ or ‘oppression’, and this construct might be manifested in the ways in which women experience and interpret violence.

To summarize, we want to make sure the people we are interviewing or the situations we are observing are representative of the targeted issues. We want to gain as complete a picture of the issues from the perspectives of the participants as possible. The usual approach to qualitative sampling is to interview or observe people in situations that fit the criteria until we have reached a point where little or no new observations are being made. This is called *saturation*.

External Validity and Sampling

The term '*external validity*' refers to the extent to which the results of an investigation can be accurately generalized to other samples or situations. External validity can be classified into two types: population and ecological.⁸

Population validity refers to generalizing the findings from the sample to the population from which it was drawn. We have already examined the importance of having a representative sample in the generalizing of results from a sample to a population. However, an investigator in the health sciences might face another problem: the accessible population from which the sample was taken might not be the same as the target population; that is, the one of general interest. Let us illustrate this point with an example.

A physiotherapist working in a large private maternity hospital intends to examine the effectiveness of a new antenatal exercise procedure for pregnant women for controlling levels of pain during delivery. A random sample of 50 pregnant women is chosen for the investigation from the population attending the hospital. The sample is then randomly assigned into two groups: one receiving the new antenatal exercise procedure, the other receiving the traditional program. The researcher finds a statistically and clinically significant difference between the two procedures, such that the new program is shown to be effective.

Strictly speaking, these findings can be generalized only to the population of women who attend the hospital. If the target population is *all* women who are having babies, then the generalization will lack external validity, because women attending different hospitals or having children at home had no chance of being included in the study sample. They might

have different characteristics, and these different characteristics may interact with the treatment in different ways to that of the sample. For instance, women who choose to deliver at home might respond better than those who choose to go to hospital, but we would not know this from this study.

Ecological Validity

There is also another facet of external validity: the situation in which an investigation is carried out might not be generalizable to other situations. This is called *ecological validity*. Consider the following examples:

1. It was shown in a classic study that morphine is an excellent analgesic for clinical pain (due to disease or injury). However, in laboratory studies of pain induced by electric shocks, morphine had little effect on participants' reports of pain threshold. Clearly, generalizing from laboratory to clinical settings, or vice versa, has to be done with extreme caution. ⁹
2. Coronary arteriography involves the insertion of a small-gauge catheter into a coronary artery and injecting a dye for x-ray visualization. It was initially reported that the mortality rates for this rather dangerous-sounding practice were only 0.1% (1 per 1000) in a first-class medical institution. However, later reports from various other institutions showed mortality rates as high as 8% (80 per 1000). ¹⁰ Clearly, the effectiveness of treatments or the usefulness of clinical evaluations can well depend on who implements them.

These two examples illustrate the caution necessary in generalizing findings.

Summary

Appropriate sampling strategies ensure the external validity of studies and their findings. The aim of sampling strategies is to ensure the selection of a sample that is *representative* of the population of objects, persons or events

the investigator aims to study. Incidental and quota samples are chosen for convenience, but these sampling strategies do not guarantee a representative sample. Random and stratified random sampling methods ensure that all important groups and characteristics in the population have the best chance to be selected and included in the study sample. Random sampling strategies are the most desirable to obtain a representative sample, although random sampling is not always feasible or desirable in health research. Area sampling and systematic sampling are also strategies that can be used to obtain representative samples.

An adequate sample size reduces the chance of large sampling error. The probable sampling error is inversely proportional to the square root of the sample size. It was argued that optimum sample size is not the maximum number of obtainable participants or a constant number or proportion. Rather, it has to be estimated for a specific investigation on the basis of the parameters of the phenomenon being studied and the study circumstances. Two types of external validity were discussed: population and ecological. External validity is related to inference, which involves using evidence from a limited set of elements to formulate general propositions.

Various approaches to purposive sampling in qualitative research were reviewed and an example of their application to a hypothetical study situation was presented.

References

1. Ianssek R, Morris M.E, eds. *Rehabilitation in Movement Disorders* . Cambridge: Cambridge University Press; 2013 doi: [10.1017/CBO9781139012942](https://doi.org/10.1017/CBO9781139012942).
2. Jones M.P, Attia J.R. Sampling: how you choose people is as important as how you analyse their data. *Med J Aust* . 2017;206(2):67–68.
3. Dal Grande E, Chittleborough C.R, Campostrini S, Taylor A.W . Bias of health estimates obtained from chronic disease and risk factor surveillance systems using telephone

population surveys in Australia: results from a representative face-to-face survey in Australia from 2010 to 2013. *BMC Med Res Methodol* . 2016;16:44. doi: [10.1186/s12874-016-0145-z](https://doi.org/10.1186/s12874-016-0145-z).

4. Gibson D.G, Pereira A, Farrenkopf B.A, Labrique A.B, Pariyo G.W, Hyder A.A. Mobile phone surveys for collecting population-level estimates in low- and middle-income countries: a literature review. *J Med Internet Res* . 2017;19(5):e139. doi: [10.2196/jmir.7428](https://doi.org/10.2196/jmir.7428).
5. Florey C.D. Sample size for beginners. *BMJ* . 1993;306(6886):1181–1184. doi: [10.1136/bmj.306.6886.1181](https://doi.org/10.1136/bmj.306.6886.1181).
6. Patton M.Q. *Qualitative Evaluation and Research Methods* . 2nd ed. Newbury Park, CA: Sage; 1990.
7. Rice P.L, Ezzy D. *Qualitative Research Methods: A Health Focus* . Oxford: Oxford University Press; 1999.
8. Huck S.W, Cormier W.H, Bounds W.G. *Reading Statistics and Research* . New York: Harper & Row; 1974.
9. Beecher H.K. *Measurement of Subjective Responses: Quantitative Effects of Drugs* . Oxford: Oxford University Press; 1959.
10. Taylor R. *Medicine Out of Control: The Anatomy of Malignant Technology* . Melbourne: Sun Books; 1979.

Ethics

Introduction,
Philosophical Principles,
 A Short History of Modern Health Research
Ethics,
 Key Principles and Concepts,
Resources,
 Balancing Participants' Rights with Research
 Benefits,
 Ethics Committees,
Summary,

Introduction

The term '*ethics*' is derived from *ethos*, referring to the values and customs of a culture at a particular point in time. Ethics are systems of moral principles that guide human action. In health we are concerned with questions such as 'How should I treat other people?' or 'What research is permissible and how should I conduct it?' In health research, where the well-being of participants is at stake, ethical considerations play a key role in research planning and implementation.

This chapter deals with the issue of ethics in the context of health research. Research that is conducted within clinics, hospitals, research

institutes and universities requires approval from an ethics committee. A research project is deemed 'ethical' if it has been approved by an ethics committee that evaluates the research questions, design and implementation of a proposed study. Sometimes the design of a study may require amendments or even may be denied approval because of ethical considerations raised by the committee.

In this chapter, we will address the values and principles which guide the way ethics review processes are conducted. Additionally, we will discuss the principles that guide the preparation of submissions and the decisions of ethics committees, and examine the basic ethical principles which govern health research.

Philosophical Principles

Ethics is a branch of philosophy concerned with the values and reasoning underlying what constitutes acceptable human conduct. Ethics has a long history with traditions traceable to ideas arising in ancient Greece and Judeo-Christian and Asian and Chinese moral traditions. Four thousand years of debate by philosophers, theologians and more recently bioethicists have ensured many different interpretations of the principles of ethics.¹ Two branches of ethics that most directly apply to conducting health research are medical ethics and bioethics. *Medical ethics* is defined as the moral conduct and principles that govern practices of medical and health professionals. *Bioethics* refers to the study of social and moral issues raised by developments in biology in general and in particular biomedicine and the health sciences. The two philosophic frameworks which are most relevant for conducting health care research are *deontology* and *utilitarianism*.²

Deontology emphasizes that individuals should be treated as an end in themselves, rather than as a means to an end. Deontological ethics hold that researchers have more duties and obligations to protect the safety of their research participants. From this perspective, research can only be justified insofar as it benefits the individuals who participate. Some philosophers believe that there are absolute standards of right and wrong which are independent of time and place. Emmanuel Kant promulgated this notion of the 'categorical imperative', which represents right and wrong for all

humans. The ancient precept, ‘above all, do no harm’ (frequently attributed to Hippocrates, but some historical research questions this; see Smith’s article on Origin and Uses of *Primum Non Nocere*—Above All, Do No Harm!), can be seen as an *absolute* standard for health practitioners and researchers.³

Utilitarianism holds that we should produce the greatest good for the greatest number of people. This is a consequentialist approach to ethics, that is, it is the outcome of our actions which define our understanding of right and wrong. The benefits of a research project should outweigh potential harm to the research participants. Utilitarianism entails a *relativist* approach to ethics, holding that there are no absolute, universally accepted principles which apply to any given historical or cultural context. Rather, ethics committees need to weigh the risks and benefits for the participants and the community associated with the research proposal.

Philosophic principles are essential; however, they do not, in themselves, enable us to determine whether health research can be judged as ‘ethical’. Codes of practice are explicit and applicable statements which are directly relevant to conducting ethical research.

A Short History of Modern Health Research Ethics

The 1964 Declaration of Helsinki by the World Medical Association was a seminal global event in the development of research ethics principles and processes. The Declaration was an elaboration and development of the 1947 Nuremberg Code.⁴ The Code was a response from the judges at the Nuremberg War Crime Trials to the cruel and harmful research projects that had been undertaken in World War II. It specifies basic principles that should govern the conduct of medical research. Most contemporary research ethics codes across the world draw heavily from the Helsinki Declaration and the underpinning Nuremberg Code.

The Declaration has been elaborated a further eight times by the World Medical Association at various meetings since 1964, with the most recent changes being in 2008. The amendments have related to drafting improvements and changes in technology and knowledge. In particular, the

responsibilities relating to publication and dissemination of research have now been specified.

Although the Declaration is now quite lengthy, it is worthwhile to reproduce the basic principles within the declaration as they provide a clear and definitive guide to the ethics principles that drive ethics processes globally. The language in the guidelines is oriented towards physicians, as it is, after all, a statement made by the World Medical Association, but the content is directly applicable to all research involving humans, and it has been used extensively in the design of human research ethics procedures more broadly. The Declaration is the foundation for most ethics statements and procedures throughout the world.

There have been numerous codes of medical ethics published in various countries and professional bodies for guiding health and medical research.⁵ The Helsinki Declaration has been very influential in the development of ethics codes in different countries. Thus although the codes vary somewhat from country to country, they have essentially the same foundation. The Declaration is built upon certain key and basic ethics principles and concepts. These principles and concepts are discussed here.

Key Principles and Concepts

There are many principles and concepts in ethics processes but some are fundamental. We will now discuss some of the basic ones. Most countries have issued statements and regulations on health and medical research ethics. The UK Medical Research Council has published a useful series of documents explaining its ethics procedures and principles (see <https://mrc.ukri.org/publications/browse/good-research-practice-principles-and-guidelines/>) as has the National Health and Medical Research Council in Australia (see the NHMRC National Statement on ethical conduct in Human Research found at <https://www.nhmrc.gov.au/guidelines-publications/e72>). Many of the guidelines share the conceptual foundations outlined in the Helsinki Declaration.

A key ethics principle is that of *informed consent and self-determination* (i.e. freedom of choice to participate or not). That is, research participants must be fully informed about the purposes of the research, any risks

associated with their participation and the uses to which the collected research data will be put. It is important that all foreseeable uses of the data are specified at the time of consent. Their participation must be fully voluntary and made with full knowledge of any potential benefits and costs. Special arrangements are implemented where the researchers may have undue influence upon the research participants such as if they are also their patients or have some form of dependent relationship with the researcher.

As discussed later in this chapter, this principle is frequently protected through the use of explanatory statements sometimes called *plain language statements*. In this statement, the researcher explains in plain language to the participants the goals of the research and exactly what is going to happen to them if they agree to participate in the research. The participant's agreement to participate with full information and knowledge about the study is recorded in an informed consent document they would normally sign. In this document, they state they have understood the purpose of the study, its risks and the nature of their participation in the research.

Another key ethics principle is *scientific excellence and quality*. Most research ethics committees take the view that the proposed research has to be of high scientific quality and address research questions that have not already been answered. This is safeguarded through the ethics review process where the evaluators consider whether the researchers have the qualifications necessary to conduct the work and whether the scientific rationale for the work is sound. The view is taken that it is unethical to conduct work that is not going to add to the scientific body of knowledge because of its poor quality or a high degree of overlap with existing work. This is a waste of resources and involves participants in unnecessary activities and potential risks.

Another key ethics principle is *minimizing risk and harm to participants*. Of course, most scientific work and indeed all human activities involve some degree of risk and the possibility of harm, but this needs to be minimized to the maximum possible extent. The Helsinki Declaration and all ethics processes that derive from it centrally consider the risk of harm to the participant. They seek to see that all risks have been minimized and that the participant is fully aware of any such risks before they have agreed to participate. In animal research, discussed later in this chapter, minimization of harm to animal research subjects also is a high priority in ethical review.

A further key ethics principle is that of *confidentiality and right to privacy*. This principle is concerned with ensuring that people's private information is not disclosed as a result of their participation in the study. In most ethics processes, this involves quite a lot of attention concerning how people's information is to be obtained, stored, processed and disseminated. The right to privacy has been particularly controversial in social media platforms in recent times. In 2018, Facebook attracted considerable attention because of unplanned third-party use of personal data by its partners. In Australia, the Digital Health Agency has been subjected to close scrutiny because of the use of its data by its commercial partners. In the United Kingdom, many agencies now have a senior officer with the title of Data Protection Officer. This person is charged with the responsibility of ensuring all laws and regulations governing privacy of personal data are followed. So in the ethics approval process, the researcher often is asked detailed questions about how the information is to be obtained and handled and how it is to be retained and ultimately disposed of when the project is completed. Further detail is required on how the work is to be published and disseminated. Many countries also now have privacy legislation that informs the local ethics procedures because any work within that country needs to comply with the local laws.

The risks of identifying individuals in research are increased in the study of small, specialized subpopulations and in qualitative studies where direct quotation of the words of the research participant may be used in publications. One of the authors of this text was once asked to participate in a study of staff who had left a university with the assurance that the results would be anonymous. However, he was the only male professor who had left over a 5-year period in his faculty. With any basic cross-tabulation of the data based on appointment level and gender, he would be the only respondent in that category, so the chances of being identified would be very high. In qualitative studies involving the direct reporting of the statements of participants, the risk is also increased. People have distinctive manners of expression and experiences such that, although there may be no names mentioned, people who know them may be able to identify the person through their language and the experiences they are describing. We know of numerous examples where others have identified research participants when their responses have been directly quoted in research

reports and publications with the promised protection of anonymity. So, given that most researchers promise to maintain anonymity of their research participants, the possibility of ‘outing’ through publication must be taken seriously and guarded against.

Often the terms ‘*confidentiality*’ and ‘*anonymity*’ are used interchangeably. However, they actually mean different things. Anonymity means that the identity of the research participant is not disclosed and in some cases may not even be known to the researchers. Confidentiality, however, means that whatever is being maintained as confidential is not revealed to a third party. Although some researchers promise confidentiality of the research findings and the participants, what they often really mean is anonymity. A literal interpretation of confidentiality would mean that they could not publish their research results!

A further key ethical principle is *lack of conflict of interest*. Conflict of interest occurs when an individual or an agency is potentially able to exploit their involvement in research in some way for their own personal or agency benefit. Some commentators argue that all researchers have an intrinsic potential conflict of interest in that they are directly affected by the findings of their research in some way or another. For example, the outcomes for the researcher may be quite different if they identify that a new therapy is effective versus if it does not work. The positive result may mean more support for their work, professional acclaim, press coverage and so on, but a negative or null result may result in defunding and little prospect of acclaim or favourable publicity. Similarly, a government agency or institution may have a vested interest in showing that a disease or problem within the community is low or even high to legitimize their existence.

The relationship between researchers and research funding bodies is also an area of potential conflict of interest. This is why many journals and professional and scientific bodies require full disclosure of such arrangements. The researchers are required to disclose their funding sources, memberships and relationships that may affect their independence and perceived bias. In ethics review, a high standard is imposed upon researchers. Thus for many, the appearance of potential conflict of interest is just as important as whether it is an actual conflict. Some researchers, for example, will not accept funding from tobacco companies, or alcohol or gambling interests, or pharmaceutical companies or government because of

a perceived conflict of interest for them. There is always the fear that the results obtained in such research may be biased in some way, either through direct influence or a desire on the part of the researchers to please the funding agency. To address this issue some of these funding agencies have 'arm's length' funding arrangements where they donate to a third-party foundation, but some are intricately involved in the approval and conduct of the research. Although government is generally not categorized with tobacco interests, government officials and ministers still have identifiable conflicts of interest and through 'contract research' can potentially exert considerable influence on the conduct and outcomes of the research they fund.

In some instances the pressure upon researchers to produce exciting and fundable research findings results in misconduct whereby the researcher may cut corners through use of low-cost or nonstandard protocols or even may directly falsify their research findings. There have been a series of celebrated examples in history and modern day of such lapses. Sir Cyril Burt is one such case. It is claimed by some that Burt falsified some key data sets for his research (a claim hotly contested to this day). Typing the search terms 'medical research fraud' into any Internet search engine brings up hundreds of reported cases of misconduct and fraud, often involving data falsification. However, it must be remembered that this is a very small proportion of the studies that are conducted. For example, in the United Kingdom it is typical to exceed more than 2 million animal study notifications annually. Interestingly, some methodologists claim that poorly conducted research and researcher incompetence, notwithstanding ethics approval, is a much more serious threat to science than outright fraud!

Finally, a key principle in ethics protocols is *independent review*. Researchers and funding bodies are not in a good position to independently evaluate the ethical implications of their own work or the work they may wish to fund. They have an intrinsic conflict of interest. This is why most medical research councils and funding bodies, whereas enunciating the principles and processes for ethical review, do not directly participate in such review. Of course, the independence and expertise of reviewers is in itself an issue in terms of potential conflicts of interest. Expert reviewers in many fields may have friendships or enmities with the person or research group whose work they may be reviewing. Even if they don't know them,

some professional jealousies may exist. This is why there are particularly detailed guidelines for reviewers to attempt to address these issues. In small, specialized fields, obtaining an independent review is particularly problematic because ‘everyone knows everyone’. A second option is to use reviewers who do not work in the field. However, the problem then is that it is easy to obtain reviews that are misinformed about important technical details of the research.

Resources

A fundamental value is that the time and effort of researchers and participants and the resources of the community should not be wasted on a badly planned investigation. Of course, there can never be a guarantee that an investigation will produce clinically useful results. However, some poorly designed projects can be assessed as doomed to fail even before the data collection begins, and they lead to confusion and controversy in the professional literature. Therefore the appropriate use of high-quality and sound research methods is not only useful for solving problems but also constitutes an ethical necessity. Many ethics committees also vet projects for scientific merit before the project is approved.

The resources available to investigators also can influence selection of appropriate research strategies. Some projects involving the evaluation of the safety and effectiveness of new drugs, or the identification of risk factors in cardiac disease, have taken tens of millions of dollars to finance. Research planning takes into account economic issues, such as:

1. *Availability of participants.* Here the investigator has to consider if enough people volunteer to participate in the project under the required conditions. Attention has to be paid to issues such as the frequency of the disorder in the community, problems in identifying participants and their level of voluntary participation. As discussed in [Chapter 5](#), selection of participants is crucial for the validity of the evidence produced by a study.
2. *Availability of equipment.* Some equipment is costly to acquire or to operate (e.g. computed tomography (CT) scanners, biochemical

assays, experimental drugs). Research planning should take these expenses into account.

3. *Availability of expertise.* Specialized assessment techniques and the administration of clinical treatments require professional expertise. If these are beyond the investigator's competence or if the experimental design requires an unbiased, 'blind' therapist or assessor (see [Ch. 7](#)), there must be means for securing them.
4. *Availability of time.* Research projects have a tendency to take considerably longer than an inexperienced researcher might expect, because of the erratic and at times disastrous workings of Murphy's Law: 'If anything can go wrong, it will'. Equipment breaks down and needs to be repaired, participants 'disappear' and collaborators might not deliver their services as promised. Research planning should take into account possible problems and how these might affect the time scale of investigations. This is a particularly important issue for postgraduate students who have limited time to finish their project.

The availability of resources strongly influences the scope of the research program and also the research strategy selected by the investigator. A research project must be shown to be economically feasible before it is initiated. It is only after the ethical constraints and economic resources have been evaluated that the researchers' questions can be transformed into clearly defined hypotheses or aims. The truth of the matter is that even if a researcher has brilliant, groundbreaking ideas, the project to gather evidence to confirm these ideas will not proceed without adequate resources.

Balancing Participants' Rights with Research Benefits

Who will benefit from the research? One likes to think that 'humanity', 'the participants' or 'health science' are going to be the beneficiaries of health research. A slightly more cynical analysis points to the investigators as having the most to gain, at least in the short term. After all, given a successful outcome for the research project, the investigators stand to

satisfy their curiosity, improve their career prospects and raise their esteem before colleagues. In practice, benefits accruing to the investigator, the participants and health science have to be carefully weighed in relation to the conduct of any project. The protection of the rights and welfare of the participants is the primary consideration of research ethics.

Sometimes harm may be caused by denying participants access to beneficial treatment as part of a study. This can happen when we use 'no treatment', 'placebo' or 'conventional treatments' as controls. This is particularly a problem when a treatment study begins to demonstrate that one treatment is clearly better than another. Many studies now set predetermined criteria to terminate the study if this occurs, thus allowing those who are not receiving a beneficial treatment to do so. Similarly, the reverse is done when a treatment is shown to be harmful.

Ethical research should ensure that the given research methodology minimizes the intrusiveness, distress and potential harm participants may experience. Distress or harm may be caused by research procedures such as the administration of a new assessment or treatment technique. It may be caused also by the strategy the researcher uses to collect information. For example, the use of existing health care records may be less distressing than reassessing patients or clients for a research study.

A range of methods has been developed to reduce the intrusiveness of research into people's lives. These include researchers working together with people in the community on tasks and projects where the participants take major roles in all the stages of conducting the research. Also there are various 'participant observer' methods available where the researcher can unobtrusively become part of the research setting, while collecting the evidence. These strategies are designed to encourage a partnership between the researcher and the participants in the research. From this perspective, the very idea of 'the research subject' is problematic since it implies that the people who participate in research are passive and have research done to them or on them.

It is important, therefore, that research is open to scrutiny by independent reviewers while it is being conducted and that research findings are replicated independently by other researchers before they are generally accepted.

Ethics Committees

Because of the large administrative burden associated with ethics review and the issue of conflict of interest, most health and medical research bodies operate under an institutional review board or research ethics committee arrangements. That is, the process of ethics review is conducted within an organization with a specially constituted group consisting of staff from the institution and external members. Most institutions that grant funds to researchers also have requirements for the constitution and operation of the institutional review boards to ensure that the reviews are high quality and conducted without conflict of interest. Specified members frequently include lawyers, ministers of religion, ethics experts and lay members, so the committees are by no means stacked with fellow researchers or people from the institution. In the event of a conflict of interest, members do not participate in the review in which they have the conflict.

If you are working in a small clinical agency or even in a large clinical agency in a developing country, it is possible that you will not have direct access to your own institutional review board. If this is the case and you are conducting some research, then you will need to gain access to a properly constituted committee to review your work. This will provide protection to you but also it means that you will be able to publish your work. The ethics review is now required for most credible research journals (i.e. you cannot publish your paper without providing evidence that you have obtained approval – generally the ethic agency and approval number and data). Most universities also require proof of ethics approval before the thesis can be submitted for publication. It is very difficult to get this after the event, so ethics approval is a key part of research planning. You will be asked to provide details of your ethics approval to have the work accepted for publication or submission for examination. So if the researcher does not have immediate access to a properly constituted and approved ethics review committee, the researcher will need to find one to review their work if they wish to publish it in a reputable journal. Some agencies charge for this service but some do not.

Recently there has been a concerning trend where agencies have attempted to avoid ethics requirements by declaring that the work is not ‘research’, that it is instead a quality improvement activity and hence ethics

approval is not required. The reasoning behind this sleight of hand is puzzling. Does this mean that it is being claimed that in quality improvement we do not need to behave according to ethical principles? Our view is that ethics procedures should be followed in all research and evaluation activities involving the collection of data from or about people.

Summary

A responsible investigator is required to take ethical principles into consideration and to plan research projects accordingly so that the probable harm to participants is minimized. Therefore the health researcher must use considerable ingenuity in designing valid investigations, while maintaining ethical values. One of the roles of ethics committees is to guide investigators on complex issues and to ensure that research is conducted in accordance with accepted community principles.

This chapter has introduced ethics principles and procedures. A short history of the development of ethics in research has been provided noting the development of the Nuremberg Code from the World War II war crimes trials concerning medical research abuses, and the World Medical Association Helsinki Declaration. The Declaration has been updated to reflect contemporary circumstances from its original to the present day. A discussion of the basic principles underlying contemporary ethics thinking and review has been provided, including informed consent and freedom of choice to participate in the research or not; scientific excellence and quality; minimization of risk and harm to participants; lack of conflict of interest; and independent review. The process of ethics review and the documentation required has been described and some tips for effective ethics submissions have been provided.

References

1. Fara P. *Science: A Four Thousand Year History*. New York, NY: Oxford University Press; 2009.
2. Duncan P. *Values, Ethics and Healthcare*. London: Sage; 2010.

3. Smith C.M. Origin and uses of primum non nocere—above all, do no harm!. *J Clin Pharmacol* . 2013;45(4):371–377.
4. World Medical Association. *World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects* . Edinburgh: 52nd WMA General Assembly; 2008.
5. Emanuel E.J, Wendler D, Grady C. What makes clinical research ethical? *JAMA* . 2000;283(20):2701–2711.

SECTION 3

Research Designs

OUTLINE

Introduction

7. Experimental Designs and Randomized Controlled Trials

8. Surveys and Quasi-Experimental Designs

9. Qualitative Research

Introduction

The aim of Section 3 is to examine various research designs and outline their applications for conducting research. As in other areas of creative activities, such as architecture or the fashion industry, the term ‘*design*’ refers to explicit plans for completing a particular activity.

The conceptual basis for experimental designs is the need for control as outlined in [Chapter 7](#). In experimental research we manipulate one or more variables (independent variables) while controlling extraneous variables through the use of appropriate control groups. If an experiment is well designed and properly conducted we are in a position to demonstrate the causal effects of independent variables on the outcome or dependent variables. The random assignment of participants to treatment or control groups is a common way to achieve control in applied health sciences research. Randomized controlled trials are experimental designs which are currently the preferred (gold standard) designs for evaluating the efficacy of interventions. In [Chapter 7](#), we will examine how the use of randomized controlled trials enables researchers to control for the biases of both the participants and the observers who evaluate the outcomes thus ensuring the internal validity of trials addressing treatment efficacy.

There are settings where for practical or ethical reasons we cannot randomly assign participants to control or treatment groups. In [Chapter 8](#), we will discuss the use of naturalistic comparisons, correlational designs, quasi-experiments and descriptive surveys. These designs enable us to describe and predict relationships among health-related variables such as risk factors influencing health. The accuracy of the information collected using these nonexperimental designs depends on the use of appropriate sampling strategies for selecting the participants used for comparisons.

In [Chapter 9](#), we will describe elementary characteristics of qualitative research designs. The aim of this chapter is to examine the many ways in

which qualitative research can be used to understand the experiences and actions of research participants. This data enables us to see health care problems and the outcomes of interventions from the perspective of the patients/participants.

Experimental Designs and Randomized Controlled Trials

Introduction,
The Concept of Causality,
Confounding and Bias in Research Studies,
The Use of Control Groups in Applied Health Research,
Assignment of Participants into Groups,
 Independent Groups,
 Matched Groups,
Types of Experimental Designs,
 Post-Test-Only Design,
 Pre-Test/Post-Test Design,
 Repeated Measures,
 Factorial Designs,
 Multiple Dependent Variables,
Randomized Controlled Trials,
Controlled Research Involving Human Participants,
The External Validity of Experiments and RCTs,
Summary,

Introduction

There is a fundamental need to identify the causes of illness and disability if we are to deliver effective health care for individuals and the population. It is by understanding these causes that we can formulate and justify our assessments, diagnoses and interventions. Furthermore, we can justify interventions in that we can point to the evidence that demonstrates that interventions are *causing* the beneficial changes in preventing or managing disease. In health research, the concept of internal validity is related to the design of research projects and the extent to which the implementations of these designs enable researchers to unambiguously identify causal relationships between interventions and outcomes.

Experiments are an important form of intervention studies. A well-designed experiment enables researchers to demonstrate how manipulating one set of variables (the *independent* variables) produces systematic changes in another set of variables (the outcome or *dependent* variables). Experimental designs aim to ensure that all variables other than the independent variable(s) are controlled; that is, that there are no uncontrolled extraneous variables or factors that might systematically influence changes in the study's outcome variable(s). Control is most readily exercised in the sheltered environment of research settings such as laboratories, and this is one of the reasons why these settings are the preferred habitats of experimental research scientists. However, much of epidemiologic and clinical research takes place in field settings (e.g. communities, hospitals and clinics), where the phenomena of human health, illness and treatment are naturally located. Even in these natural settings, researchers can nevertheless exercise control over extraneous variables.

In this chapter, we will focus on an experimental design referred to as the *randomized controlled trial* (RCT), which is an experimental research design aimed at assessing the effectiveness of a clinical intervention. Our aim is to show how RCTs are used to demonstrate that an intervention is the cause of a subsequent effect(s). RCTs are considered by many researchers and clinicians to be the premier type of evidence for the effectiveness of clinical interventions.

The specific aims of this chapter are to:

1. Examine the basic structure of experimental research designs and RCTs.
2. Consider threats to the validity of the results obtained from experiments and RCTs.

The Concept of Causality

The notion of causality has been a difficult and controversial topic for philosophers. As health researchers, we need to take a pragmatic approach and look to demonstrate causal relationships. There are three simple criteria for demonstrating causal associations.

1. The cause must precede (occur before) the effect.
2. The cause and effect co-vary. If the cause occurs, then so does the effect.
3. If the cause does not occur, then the effect does not occur.

The first criterion is quite simple. For example, if we say that injury to the person's arm is the cause of that person's reported pain, we of course assume that the injury was sustained before the onset of the pain. Clearly, if the pain was already present, the injury would not be seen as the cause of the pain. Second, we assume that there will be 'concomitant variation' between the injury and the pain. The worse the injury, the more severe the pain. As the injury recovers, a decrease in the level of pain can be expected. In general, we are establishing evidence for the existence of a causal relationship between the cause and the effect.

However, observing a relationship between two events is not sufficient to demonstrate causality. For example, night follows day in a predictable, lawful fashion, but we do not say that day causes the night or vice versa. We must attempt to eliminate plausible alternative explanations or hypotheses that offer rival causal explanations for the findings. For example, pain might persist even after the injury has healed. There might be other variables operating which maintain a person's experience of pain.

The problem is that apart from our intervention, there are likely to be other factors that may be influencing the pain outcome. These influences are called 'confounding extraneous variables' or 'bias' and constitute a

potentially serious problem when attempting to interpret the results of a trial.

Confounding and Bias in Research Studies

The research designs that are commonly referred to as ‘before–after’ designs often are used initially to test the safety and efficacy of novel interventions. The use of the design provides a clear example of how confounding variables can threaten the internal validity (or accuracy) of the study. Confounding variables can result in ambiguity in deciding what has caused any observed changes.

As a hypothetical example, consider the introduction of an exercise program for cardiac patients which aimed at increasing mobility and thereby improving health. The patients were also smokers and were strongly encouraged to give up smoking. The researchers used ‘distance walked by patients’ (in a specified time period) as an indicator of the effectiveness of the program. [Fig. 7.1](#) represents (without showing numbers) the average walking distances before and after the exercise program. There is clearly a difference between before and after (i.e. an improvement from baseline). However, was this improvement caused only by the exercise program, or by confounding extraneous variables, or a combination of both?

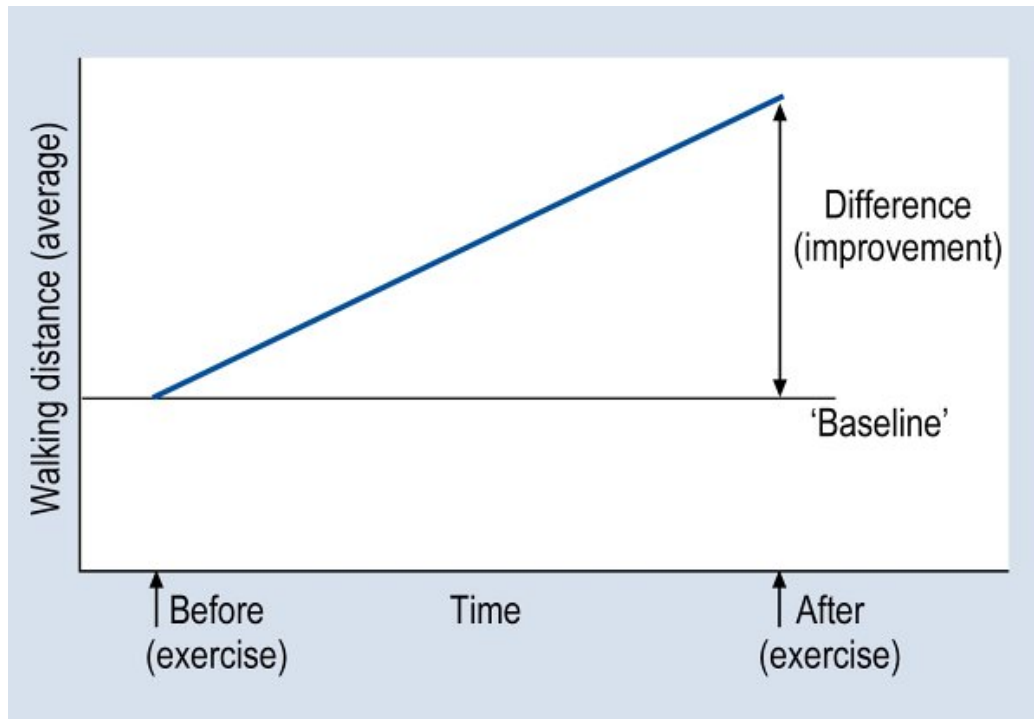


FIG. 7.1 Results for a hypothetical 'before–after' study.

Consider the these four plausible alternative explanations for the difference shown in [Fig. 7.1](#):

1. The improvement might have been due to natural recovery; the patients' mobility might have improved independently of the exercise program. This is referred to as a threat to internal validity due to *maturation*.
2. The improvement might have been due to other factors, such as the reduction in, or cessation of, smoking among some of the patients, resulting in an average increase in walking distance. Such confounding extraneous variables are referred to as *history*.
3. The improvement might have been due to a placebo effect. Placebo effects (see later) are associated with patient expectations of potential benefits of the intervention. In the example outlined previously, the expectations of the patients, rather than the exercise itself, might have been the actual cause of the improvement.
4. An important source of bias is *observer bias*, which refers to researchers unknowingly influencing the results of a study by

holding expectations regarding the outcome. In the present hypothetical example, the researchers may anticipate that the exercise program will improve mobility.

These explanations illustrate the influence confounding extraneous variables. That is, each is concerned with factors that are present at the same time as the intervention and which may have produced the observed effect.

Internal validity refers to the ability of a researcher to attribute differences (e.g. [Fig. 7.1](#)) to the effect of the independent variable. Maximizing internal validity is important for quantitative research designed to demonstrate causal effects. To ensure internal validity, researchers must control for the effects of confounding and bias. Experimental designs are particularly well suited to controlling for the potential influence of confounding factors. RCTs employ experimental designs that are applied for evaluating the efficacy of interventions.

The Use of Control Groups in Applied Health Research

A control group consists of participants who undergo the same conditions as the group receiving the intervention under investigation. For example, in drug trials, control group participants often will receive an injection of saline solution, if the experimental treatment is administered via injection, to control for the effects of the actual injection. If the medication were administered orally, similar-looking inert tablets would be used for control participants. It has been found that if people receive any form of ‘therapy’, improvement may occur even when the ‘treatment’ or intervention is physiologically and chemically inert. This is referred to as a *placebo effect*. The control group allows the researcher to measure the size of the placebo effect, and to take it into account when interpreting the study results. If we are to include a control group in our intervention studies, it is essential that, at the outset, the experimental and control groups are as similar as possible. We have to take the participants and split them up into experimental and control groups as equally as possible. This process is called *assignment*.

The assignment of participants to their groups by the investigator is an essential feature of an intervention study.

Let us reexamine the investigation outlined in [Fig. 7.1](#) in terms of the effect upon internal validity when we use a control group in the study.

1. Assess the walking distances of the patients in a pretreatment test.
2. Assign participants into experimental or control groups by matching on the basis of pre-test performance, and random assignment within pairs.
3. Administer:
 - a. Experimental intervention (exercise program, walking, diet and smoking reduction).
 - b. Control intervention (an alternative activity, walking, diet and smoking reduction).

The *only* difference between the two groups is that one group receives the exercise program and the other some alternative activity.

4. Test both groups on walking distance after the treatment.

The results of this fictional study are presented in [Fig. 7.1](#).

Let us examine how this new design stands up to threats of internal validity in contrast to the original investigation.

1. *Natural recovery.* Both the control and experimental groups had the same time to recover or deteriorate; it is unlikely that this factor explains the difference.
2. *Confounding due to other interventions.* Both groups had walking, smoking decrement and diet. It is unlikely that the difference between the post-test results shown in [Fig. 7.1](#) would be due to confounding factors.
3. *Placebo effect.* Given that this is not a double-blind trial, the bias due to expectations remains uncontrolled with this design.

We can see that using a design with a control group resulted in an improvement in the internal validity of the trial. Note that the use of a

control group has not removed the effects of the placebo. It is still possible, but much less likely, that the differences in the outcomes for the two groups were not a result of the intervention. Nevertheless, the investigator has a much sounder basis for deciding whether or not the exercise program was effective.

It should be noted that there are ways, other than the use of control groups, to minimize the effects of uncontrolled extraneous variables. For example, if ‘noise’ is a possible extraneous variable in administering a test, we may ‘insulate’ from it by using a quiet setting. However, if noisy settings are the norm in real life, then, as discussed later, the search for higher internal validity may be at the cost of external validity.

Assignment of Participants into Groups

Independent Groups

The simplest approach is to assign the participants randomly to independent groups. Each intervention group represents a ‘level’ of the independent variable. Say, for instance, we were interested in the effects of a new drug (we will call this drug A) in helping to relieve the symptoms of depression. We also decide to have a placebo control group, which involves giving patients a capsule identical to A, but not containing the active ingredient.

Given a sample size of 20, we would assign each participant randomly to either the experimental (drug A) or the control (placebo control) group. Random assignment could involve encoding all the 20 names into a computer program and using random numbers to generate two groups of 10. In this case we would end up with two equivalent groups ([Table 7.1](#)).

Table 7.1

Levels of the Independent Variables

	Control group (placebo)	Experimental group (drug A)
Number of participants	$n_c = 10$	$n_e = 10$

Here n_c and n_e refer to the number of participants in each of the groups, given that the total sample size (n) was 20. We can have more than two groups if we want. For example, if we also included another drug ‘B’, the independent variables would have three levels. We would require a total of 30 participants if the group sizes remained at 10.

Matched Groups

Random assignment does not guarantee that the two groups will be equivalent. Rather, the argument is that there is no reason why the groups should be different. Although this is true in the long run, with small sample sizes chance differences among the groups may distort the results of an experiment or RCT. Matched assignment of the participants into groups minimizes group differences caused by chance variation.

Using the hypothetical example discussed previously, say that the researcher required that the two groups should be equivalent at the start of the study on the measure of depression used in the experiment. Using a matched-groups design, the participants would be assessed for level of depression before the treatment and paired for scores from highest to lowest. Subsequently, the two participants in each pair would be assigned randomly to either the experimental group or the placebo control group. In this way, it would be likely that the two groups would have similar average pre-test depression scores.

Types of Experimental Designs

Four types of experimental design will be discussed. These are the post-test-only, pre-test/post-test, repeated measures and factorial designs.

Post-Test-Only Design

At first it may appear that this would make measurement of change impossible. At an individual level, this is certainly true. However, if we assume that the control and experimental groups were initially identical and that no change occurred in the controls, direct comparison of the post-test scores would indicate the extent of the change. This type of design is fraught with danger in clinical research and only should be used in special circumstances, such as when pre-test measures are impossible or unethical. The assumptions of initial equivalence and of no change in the control group often may not be supported and, in such cases, interpretation of group differences is difficult and ambiguous. For example, if we conduct a double-blind RCT to determine the efficacy of a drug in treating very severe infections then, if our outcome is proportion of patients surviving, a test-only design would be appropriate.

Pre-Test/Post-Test Design

In this design, measurements of the outcome or dependent variables are taken both before and after intervention. This allows the measurement of actual changes for individual cases. However, the measurement process itself may produce change, thereby introducing difficulties in the clear attribution of change to the intervention on its own. For example, in an experimental study of weight loss, simply administering a questionnaire concerning dietary habits may lead to changes in those habits by encouraging people to scrutinize their own behaviour and hence modify it. Alternatively, in measures of skill, there may be a practice effect such that, without any intervention, the performance on a second test will improve. To overcome these difficulties, many researchers turn to the post-test-only design.

Repeated Measures

To economize the number of participants required in an experimental design, researchers will sometimes reuse participants. Thus at different times the participants may receive drug A or drug B. If it were the case that every participant encountered more than one level of the drug variable or factor, then ‘drug’ would be termed a repeated measures factor. An important consideration is using a ‘counterbalanced’ design to avoid series effects. For example, half the participants should receive drug A first, and half drug B. If all the participants received drug A first and then drug B, the study would not be counterbalanced and we would not be able to determine whether the order of administration of the drugs was important. Time is a common repeated measures factor in many studies. A pre-test/post-test design involves the measurement of the same participants twice. If ‘time’ is included in the analysis of the study, then this is a repeated measures factor. In statistical analysis, repeated measures factors are treated differently from factors where each level is represented by a separate, independent group. This is true both for matched groups and for repeated measures, both discussed earlier.

Table 7.2

Examples of a Factorial Design

	Drug A	No drug
Psychological therapy	1	2
No psychological therapy	3	4

Factorial Designs

A researcher often will not be content with the manipulation of one intervention factor or variable in isolation. For example, a clinical psychologist may wish to investigate the effectiveness of both the type of psychological therapy and the use of drug therapy for a group of patients. Let us assume that the psychologist was interested in the effects of therapy versus no psychological treatment, and of drug A versus no treatment.

These two variables lead to four possible combinations of treatment ([Table 7.2](#)).

This design enables us to investigate the separate and combined effects of both independent variables on the outcome measure(s). In other words, we are looking for interactions among the two (or more) factors. If all possible combinations of the values or levels of the independent variables are included in the study, the design is said to be fully crossed. If one or more is left out, it is said to be incomplete. With an incomplete factorial design, there are some difficulties in determining the complete joint and separate effects of the independent variables.

So that the terminology in experimental designs is clear, it is instructive to consider the way in which research methodologists would typically describe the example design in [Table 7.2](#). This is a study with two independent variables (sometimes called factors): namely, type of psychological therapy and drug treatment. Each independent variable, or factor, has two levels or values. This would commonly be described as a 2×2 design (each factor having two levels). There are four groups (2×2) in the design.

If a third level, drug B, was added to the drug factor, then it would become a 2×3 design with six groups required. Two groups, drug B with psychological therapy and drug B with no psychological therapy, would be added. It is possible to overdo the number of factors and levels in experimental studies. A $4 \times 4 \times 4$ design would require 64 groups; that is a lot of research participants to find for a study. It is relevant to note that when we evaluate two or more groups over a period of time, we are also using a factorial design.

Multiple Dependent Variables

Just as it is possible to have multiple independent variables in an experimental study, it is also sometimes appropriate to have multiple outcome or dependent measures. For example, to assess the effectiveness of an intervention such as icing of an injury, factors such as extent of oedema and area of bruising are both important outcome measures. In this instance there would be two outcome measures. The use of multiple dependent variables is very common in health research. The outcomes measured are

usually evaluated individually, although there are more complex statistical techniques which enable the simultaneous analysis of multiple dependent variables.

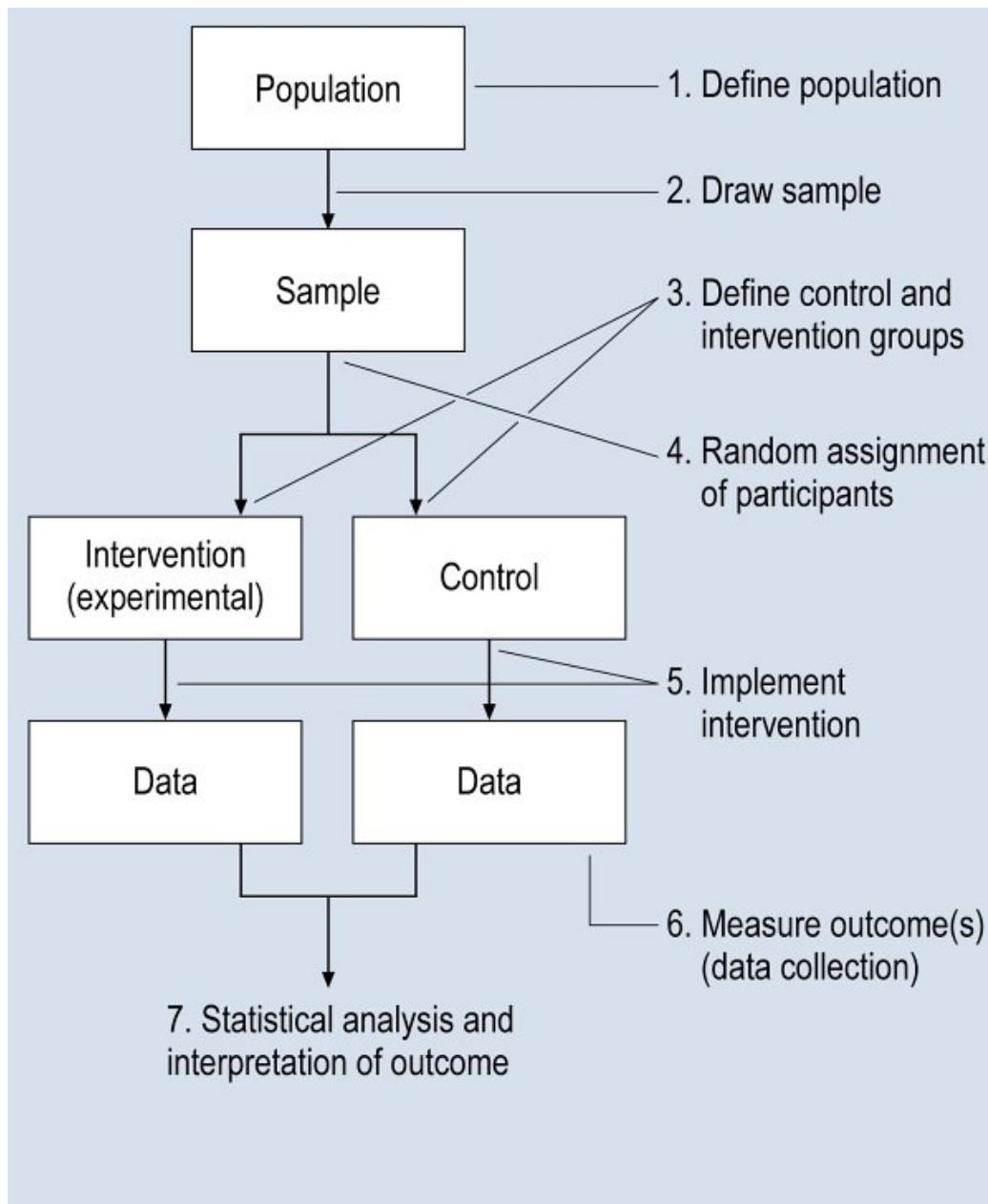


FIG. 7.2 A two independent groups randomized controlled trial.

Randomized Controlled Trials

We will now work through each of the steps in conducting an RCT; these steps are summarized in [Fig. 7.2](#). The protocol being followed has been developed by the CONSORT Committee to promote the standardization of RCTs.

1. *Definition of the population.* Researchers define the population to which they wish to generalize. For example, this might be males over 55 years of age with coronary heart disease, or the local community or a certain type of health care organization.
2. *Selection of the sample.* Using an appropriate sampling method, the study sample is selected from the population. It is desirable that the sample is representative of the population. It is important to note that steps 1 and 2 are common to all research designs (see [Ch. 5](#)).
3. *Defining the intervention and control groups.* The simplest RCT design has two groups (see [Fig. 7.2](#)): an intervention, or experimental, group and a control group. These two groups are sometimes referred to as comprising the two ‘arms’ of a study. This ‘minimalist’ design enables the researcher to compare the outcome for the intervention group with the outcome for the control group and thereby discount the effects of confounding factors. It is possible, of course, to produce far more complex designs, including those with numerous intervention groups and outcome measures. It is likely that, in the course of searching for research evidence that can be applied to practice, you will encounter more complex designs.

An essential consideration in planning an RCT is the type of control group. There are several possibilities, including:

No-intervention/waiting-list control group.

Traditional intervention control group.

Placebo control group.

The first of these (no intervention) is the most economical to implement, in the sense that little has to be done to the participants in the control group, beyond making predetermined assessments.

The problem is that there is no control for some confounding factors, such as differences in the behaviour of the participants in the two groups. Thus for example, participants in a no-intervention

control group may feel neglected and (unbeknown to the researcher) may seek alternative interventions, while remaining in the trial. Finally, no-treatment controls might be unacceptable on ethical grounds (i.e. it is usually unethical to deny treatment if it is available).

When an intervention needs to be provided to all participants, and when there are interventions already available to prevent or to treat a particular health problem, then the traditional or currently available intervention can be provided to the control group. The outcome for the new intervention group is then compared with that for the control group, principally in terms of any superiority of the new intervention.

4. *Assignment procedures.* Random assignment means that each participant has an equal chance of being assigned to any of the groups. In an RCT, one of the groups receives a control intervention with the active component of the intervention missing. For example, in an experimental study of the effectiveness of different types of weight loss programs, one group might receive an instruction manual, another might receive supervised dietary training and a third group may receive no intervention. The purpose of the randomized assignment procedure is to ensure that the groups are as similar as possible on all the relevant variables. If the groups are substantially different, it will be very difficult to attribute any differences in final outcome to the ‘treatments’ being administered.
5. *Administration of intervention (treatment).* The researcher then administers the intervention(s) to the various groups participating in the experiment. This is called the independent variable(s). It is important that the intervention is administered in an unbiased way, so that a fair test of any differences in outcomes may be provided. As will be discussed later, awareness of the expected outcomes on the part of participants may lead to a spurious biasing of the outcomes for those participants. Therefore the true aims and expectations of the researcher are sometimes concealed from the participants in an attempt to control expectancy effects.

6. *Measurement of outcomes.* The researcher assesses the outcome of the experiment through measurement of the dependent or outcome variable(s). Sometimes the dependent variable is measured both before and after the experimental intervention (this is called a pre-test/post-test design) and other times only afterwards (this is called a post-test-only design).
7. *Statistical analysis and interpretation of the outcomes.* In general, outcome measures index any changes in the health status and functioning of the participants, subsequent to the intervention, and the greater the change, the more effective the intervention is likely to be. Of course, when the intervention is implemented as part of an RCT, the intervention group(s) is always compared with the control group, so it is the difference between the intervention and control groups that represents the true effect of the intervention. Measures of the effect of the intervention are discussed briefly in the next section. The calculation of such measures and procedures for determining whether the effect is statistically or practically significant are discussed in [Chapters 20 and 21](#).

There are a number of ways of designing and implementing RCTs. An important consideration is whether the trial is explanatory or pragmatic. A pragmatic trial tests the *efficacy* of the intervention in a controlled research setting with selected participants and under ideal conditions. An efficacy study is designed to test whether the intervention works under ideal circumstances.

An effectiveness or pragmatic RCT study tests the *effectiveness* of the intervention in everyday practice. The study includes participants representative of normal patients and with less controlled conditions that more closely reflect everyday clinical practice. Pragmatic or effectiveness RCTs are used in research translation studies to check that interventions that work under highly controlled circumstances will work in normal clinic circumstances.

It is typical to conduct an efficacy study under ideal circumstances before conducting an effectiveness study in an everyday clinical setting.

Controlled Research Involving Human Participants

Investigations involving human participants require that researchers should consider both psychological and ethical issues when designing experiments. Human beings respond actively to being studied. When recruited as a research participant, a person might formulate a set of beliefs about the aims of the study and will have expectations about the outcomes. In health research, placebo effects are positive changes in signs and symptoms in people who believe that they are being offered effective treatment by health professionals. These improvements in signs and symptoms are in fact elicited by ‘inert’ treatments probably mediated by the patients’ expectations. Health professionals who hold strong beliefs about the effectiveness of their treatments communicate this attitude to their patients and possibly increase the placebo effects.

Placebo responses as such are not a problem in everyday health care. Skilled therapists use this effect to the patients’ benefit. Of course, charlatans exploit this phenomenon, masking the poor efficacy of their interventions. In health research, however, we need to demonstrate that an intervention has therapeutic effects greater than a placebo, hence the need for a placebo control group. The ideal standard for an experiment is a *double-blind* design, in which neither the research participants nor the health providers/experimenters know which participants are receiving the active form of the treatment. Placebo-controlled, double-blind RCTs, with large sample sizes, are the strongest method for controlling confounding extraneous variables, and they have a greater degree of internal validity than does any other research design. This is because the only thing that differs between the groups is whether they receive the active version of the intervention. Therefore if an effect of the active version of the intervention is demonstrated (i.e. there is a difference between groups that favours the intervention group), then we can have a good deal of certainty in the conclusion that the intervention was the cause of this effect. Placebo-controlled, double-blind RCTs are therefore sometimes referred to as constituting a ‘gold standard for research design’. However, there are situations where double-blind RCTs are not appropriate. An obvious

example is when it is impossible or very difficult to devise and to conceal a placebo intervention. For example, psychologists, speech pathologists and occupational therapists find it difficult to devise behavioural placebo interventions that can be concealed from participants. Also, as stated earlier, placebo controls might be unethical and therefore inappropriate when established treatments are available. Another source of ethical concern has been the use of placebo or 'sham surgery' for implementing double-blind RCTs of the effectiveness of some surgical interventions. It is one thing to administer inert treatments to volunteers, but another thing entirely to cut open people's bodies for the sake of creating a control group. Polgar and Ng argue that it is unethical to use placebo surgery as a means for creating a control group.¹ Rather, novel surgical procedures are best evaluated using 'traditional intervention' control groups.

It is not always possible, however, to form placebo control groups. In the area of drug research and some other physical treatments it can be possible; but with certain behaviourally based interventions, in areas such as psychotherapy, physiotherapy or occupational therapy, double-blind placebo interventions may be impossible to implement. For instance, how could a physiotherapist offering a complex and intensive exercise program be 'blind' to the treatment program being offered? In these situations, researchers employ no treatment or traditional intervention control groups when evaluating the safety and effectiveness of a novel intervention.

Also, there are ethical issues that need to be taken into account when using placebo or non-intervention control groups. Where people are suffering from acute or life-threatening conditions, assignment into a placebo or no-treatment control group could have serious consequences. This is particularly true for illnesses where prolonged participation in a placebo control group could have irreversible consequences for the sufferers. Under such circumstances, placebo or no-treatment controls might well be unethical and the selection of a traditional treatment group would be preferred.

The External Validity of Experiments And Rcts

As we discussed previously, all health research which relies on drawing representative samples may involve systematic sampling error. Sampling error or selection bias may threaten the external validity of a study, defined as the extent to which we are confident when applying the findings to a specific group or the overall population.

To consider an obvious illustration, say that you are working with a group of adults (65 years or older) with heart disease living in Montreal. You are looking for a diet program which will help optimize weight. Your literature search identifies a well-designed RCT which clearly indicates that a given diet 'X' has provided large weight losses in overweight children living in Shanghai. Would you adopt diet 'X' for your patients? Before you did, would you carefully consider the differences between your patients and the study sample, including the variables age, health status and environmental and cultural settings?

The highly structured and controlled implementation of RCTs may also influence the extent to which their findings can be applied to everyday health care situations. As an example, consider the use of double-blind RCTs to evaluate the efficacy of cellular therapies for the treatment of Parkinson's disease (see [Ch. 6](#)). The double-blind design stipulates that volunteers with Parkinson's disease are assigned to either a sham-operated control group or a treatment group in which the patients were administered human fetal cells to replace the dopaminergic neurons lost through the disease process. Therefore the concealment of the treatment means that the patient is uncertain whether she or he has actually received new cells or just a sham operation. This uncertainty is quite different to actual cellular therapy where the patient is certain of the intervention and may carry out activities for optimally integrating the new cells into their impaired neural network. That is, the trade-off between ensuring internal validity may compromise the external validity of double-blind RCTs.

Summary

The experimental (RCT) approach to research design involves the active manipulation of the independent variable(s) through the administration of an intervention often with a non-intervention control group and the measurement of outcome through the dependent variable(s). Good

experimental design requires careful sampling, assignment and measurement procedures to maximize both internal and external validity.

Common experimental designs include the pre-test/post-test, post-test-only, repeated measures and factorial approaches. These designs ensure that investigations can show causal effects. More recent approaches to evidence-based health care have emphasized the importance of employing experimental trials wherever possible.

External validity is concerned with the researcher's ability to generalize findings to other samples and settings (i.e. the generalizability of the study findings). The external validity or generalizability of a study is affected by sample size, method of sampling, and design characteristics and measures used in the study. If we say a study has high external validity, we mean that its findings generalize to other settings and samples outside the study. It does not make sense to talk about the external validity of a particular test.

Internal validity is concerned with the design characteristics of research projects. If a study is internally valid, any effects/changes or lack thereof in the dependent (outcome) variable can be directly attributed to the manipulation of the independent variable. It is important not to confuse the meanings of these terms.

Reference

1. Polgar S, Ng J. Ethics, methodology and the use of placebo controls in surgical trials. *Brain Res Bull* . 2005;67(4):290–297.

Surveys and Quasi-Experimental Designs

Introduction,
Naturalistic Non-RCT Research Designs,
Surveys,
Epidemiology,
Designs for Comparing Groups,
 Natural Comparison Studies,
 Case-Control Designs,
 Correlational Designs and Outcome Prediction
 Studies,
Quasi-Experimental Research Designs,
 Time-Series Cohort Designs,
 Multiple-Group Time-Series Designs,
The Internal and External Validity of Naturalistic Non-RCT
Designs,
Summary,

Introduction

In medical and health sciences research, the evidence-based medicine (EBM) evidence hierarchy often is used to describe the weight or confidence that should be assigned to the evidence associated with particular categories of studies.¹ Fig. 8.1 presents a typical example of the hierarchy. (There are many very similar versions of this based on the original Canadian Medical Association's Working Group work.)

Systematic reviews are highly structured evidence summaries about a particular intervention or research using rigorous selection criteria to select typically randomized controlled trials (RCTs) for analysis and inclusion. We review systematic reviews in Chapter 23 in this book. RCTs generally involve the random allocation of participants to alternative treatments and a group receiving usual treatment or placebo to compare outcomes. Some health disciplines refer to RCTs as experimental studies, but the meaning is the same. A *cohort study* is a systematic tracking of a group of people usually at periodic time intervals. Typically, two subgroups are compared: an exposed group that has been exposed to a particular clinical factor and a non-exposed comparison group. These studies can be prospective or retrospective but are usually prospective. A *case-control study* compares people who developed the targeted condition with a group that did not. A *case study* documents, generally for a single illustrative patient, the symptoms, diagnosis, treatment and outcomes. Obviously, an *expert opinion* is a balanced authoritative statement about a specific clinical situation or case.

Although the evidence hierarchy is a useful contribution to the consideration of clinical research evidence in formulating strategies for the best treatments, there has long been a commentary about the ill wisdom of uncritical acceptance of the hierarchy as a definitive guide to study quality. And indeed, tools such as GRADE (Grading of Recommendations, Assessment, Development and Evaluations; <https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/what-is-grade/>) were devised to provide a more precise assessment of study bias, based on criteria that are more detailed than a simple categorization into study type. This commentary recognizes that use of a particular study design does not guarantee that its results will be accurate and intrinsically more valuable than a study using a design that is 'lower' on the hierarchy. Indeed, it must

be remembered that in most systematic reviews, the bulk of the available RCTs are actually rejected because of high-bias/poor-quality issues in the review process. A top-quality expert opinion or editorial may provide more valuable insights than an uninspiring and poorly performed systematic review. It also should be noted that the hierarchy is primarily meant to be applied to assist in the selection of clinical interventions. Although this focus on clinical interventions is commendable, there are many health and medical research studies that are not simply about interventions and their effectiveness.

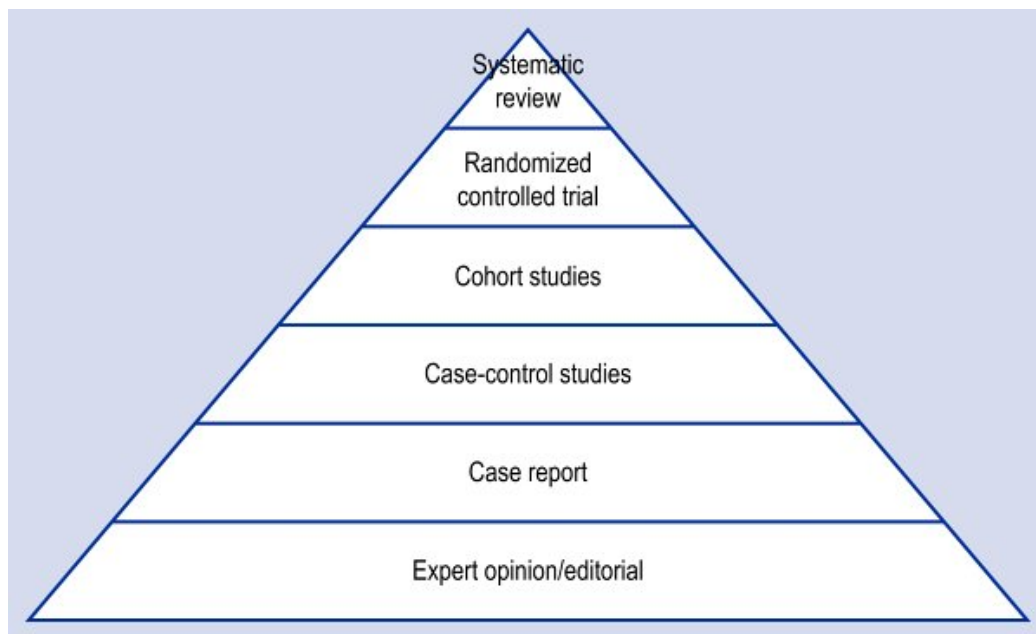


FIG. 8.1 Evidence-based medicine hierarchy

For example, epidemiologic studies are concerned with the analysis of the distribution of health and illness and their associated determinants and risk factors in defined populations. Rather than solely using randomized controlled interventions, epidemiology uses large data sets including surveys. Surveys are widely used in clinical epidemiologic research and also to collect risk and outcomes; patient experience data and also of course qualitative studies do not fall within the EBM evidence intervention hierarchy framework.

The specific aims of this chapter are to:

1. Examine the uses of naturalistic research designs.
2. Discuss the use of surveys in health sciences research.
3. Outline the characteristics of designs involving naturalistic comparisons, correlations and quasi-experiments.
4. Identify some of the limitations of these designs.

Naturalistic Non-Rct Research Designs

If RCT/experimental designs are supposed to provide the tightest possible control of extraneous factors, why should we resort to alternative research designs? There are a number of reasons why non-RCT research designs should be employed instead.

1. Many patient characteristics are not amenable to researcher manipulation (i.e. they are unchangeable). For example, if the research question is concerned with sex differences in response to heart surgery, then sex cannot be manipulated by the researcher. Similarly, if the researcher is interested in age differences, the ages of the participants cannot be altered by the researcher. But certainly the associations between these parameters and outcomes can provide highly valuable and clinically relevant data.
2. Often, it is ethically inappropriate to investigate research questions using a direct intervention design. For example, if a researcher wishes to perform a study on the effects of smoking on health, studying this in an RCT intervention study would require the researcher to allocate participants randomly to the smoking or nonsmoking group. Of course, it is unethical to force some participants to smoke and others not to smoke, so we resort to cohort or case-control methods. Furthermore, in intervention studies using a no-treatment control group, valuable and effective treatment might be withheld from participants if we were to withhold treatment. This situation involves serious ethical concerns that might lead to an experiment not being approved by the relevant ethics committee because of potential harm to the participants. It is usual to employ a comparison group that is receiving the current usual care.

3. Studies such as RCTs are best used to study simple causal relationships between variables (i.e. does the intervention cause positive therapeutic effects?). However, many human diseases and illnesses are not determined by a single cause but rather by a number of causes interacting in a complex fashion. For example, heart disease may be caused by a combination of factors such as smoking, excessive stress, inappropriate diet or genetic factors. To identify such possible causal (or risk) factors, we need to study systems as they function in nature. That is, we should investigate patients in their natural settings, even with the difficulties this entails. See the discussion of efficacy versus effectiveness studies in [Chapter 23](#) for consideration of this issue.

Depending on the nature of the evidence we require, non-RCT designs often are highly appropriate alternatives to RCT designs in intervention studies and necessary in studies with other aims. In the health and medical sciences, RCT designs are focused on a rather narrow band of research questions that relate to the effectiveness of interventions and the causes of diseases. There are many other interesting research questions that do not relate to intervention effectiveness and, even when the questions do relate to interventions, there are circumstances in which nonexperimental methods are more appropriate. Descriptive surveys are nonexperimental designs which provide data essential for progress in the clinical sciences and public health.

Surveys

Surveys are investigations aimed at describing accurately the characteristics of populations for specific variables ([Fig. 8.2](#)). Surveys are commonly used in health research for a number of purposes.

1. To establish the attitudes, opinions or beliefs of persons concerning health-related issues. The data collection techniques often include questionnaires or interviews.
2. To study characteristics of populations on health-related variables, such as use of health care, blood pressure, emotional problems or drug use patterns.

3. To collect information about the demographic characteristics (age, sex, income, etc.) of populations. A government census can be an important source of knowledge concerning population characteristics.

The statistics obtained from surveys can present an overview of the patterns of states of health, illness and the use of health services in a given community. In this way, we can gain insights into issues such as the prevalent causes of death or the health-related requirements of the population. The outcomes of the surveys can be the bases for hypotheses and theories concerning the causes of illness in a community. The area of health science concerned with such matters is called *epidemiology*.

Epidemiology

Epidemiology is the field of study that is concerned with the distribution and determinants of health and illnesses in groups of people. Epidemiology has been defined as the study of determinants of health and illness in a given population. Epidemiologic studies may be descriptive, where researchers study the incidence and prevalence of health and illness, or analytical, where researchers aim to identify the multiple and interacting factors that determine health and illness in a specific community. Epidemiologic investigations are essential to data collection in public health, as this discipline is concerned with the study of health populations. It differs from much clinical research in that it is oriented to the population or group rather than individual level. Epidemiology goes back to the time of Hippocrates who was concerned with the effects of environments upon the health of populations. The work of John Snow in the 19th century concerning the epidemiology of cholera is also considered a major landmark in the development of the discipline. Snow mapped the distribution of cholera cases in London and demonstrated that the water supply to different houses supplied by different water companies was involved with the transmission of the disease. The existence of the cholera organism was inferred from these observational data.

Descriptive epidemiology focuses on specific target populations and compares the occurrence of different health or disease states within these populations. Thus the numbers of cases, new cases and the relative risk for

having the target condition are key components of the epidemiologic approach. Two keyterms in epidemiology are '*incidence*' and '*prevalence*'. The *incidence* of a disease is the rate at which new cases occur in a specified population during a specified period. The *prevalence* of a disease is the proportion of a specified population that has the target condition at a specified point in time.

These basic definitions often are used to construct indicators of population health that can be compared across different countries. The World Health Organization (WHO) maintains statistics concerning rates such as the birth rate, the infant mortality rate and the perinatal mortality rate. The WHO website provides a wealth of comparative epidemiologic data (<http://www.who.int/en/>).

Analytical epidemiology is concerned with how diseases are transmitted or spread within populations. The classic epidemiologic approach conceptualizes disease within the framework of host, agent, vector and environment. The host is the human, the agent is the infection or health problem, the vector is the means by which the disease is carried (e.g. in the case of malaria this may be a mosquito) and the environment is the setting or mechanism that promotes the exposure. Disease can be transmitted directly (from person to person) or indirectly (via water, such as was demonstrated in Snow's classic study of London water supplies). Although this framework was developed initially within the context of infectious diseases, it is actually much more broadly applicable. Contemporary epidemiologists also study the prevalence and determinants of noncommunicable diseases (e.g. Parkinson's disease) in the population.

The pattern of outbreaks of disease within populations is also an area of interest to epidemiologists. Common patterns of occurrence of diseases include endemic, epidemic and pandemic patterns. A disease is said to be endemic to a particular area if it habitually occurs within the area. An epidemic is the occurrence of a disease that is above and beyond the usually expected rate. We often talk of flu epidemics, meaning an unexpectedly high rate of the occurrence of influenza. A pandemic is a worldwide epidemic. We have become acutely aware of pandemics such as HIV/AIDS and other feared pandemics such as severe acute respiratory syndrome, and, more recently, avian flu.

Epidemiology uses both experimental and nonexperimental designs for collecting evidence. (Interested readers should consult Bonita et al.)²

Designs for Comparing Groups

Natural Comparison Studies

Essentially, this type of study involves the comparison of naturally occurring groups or populations. For example, a researcher may be interested in the relative performances of males and females on a test of spatial abilities. The researcher might take a sample of each sex studying at a university and then compare the relative performances of the males and females on the test. Alternatively, a researcher might be interested in whether people growing up in different cultures have different pain reactions. Here, the researcher would select a sample of volunteers from the cultural groups of interest and compare their pain responses with standard pain stimuli.

There are extraneous variables which can be controlled in this type of investigation. In these studies, the researcher can control variables such as the ages or educational backgrounds of the participants by selecting similar groups. However, the researchers have not actually changed or manipulated the variables being studied. All that has been done is to make provision for the measure of differences between naturally occurring groups.

To relate this to our first example, if our researcher determines that there is indeed a statistically significant difference in the performance of males and females on the measure(s) chosen, well and good. If our researcher then claims that biological factors are solely responsible for the differences, this is another matter altogether. This inference is only one of a number of alternative explanations that could be advanced to account for the differences observed. It is possible, for instance, that males and females in a sample have been exposed to different types of toys and games during their development, so that males might perform better or worse on some tasks. In the second example, any significant results in the pain responses might not be due to ‘culture’ but rather to systematic biological differences among the groups or a complicated interaction between these factors.

The simple fact of the matter is that it is difficult to demonstrate causation in natural comparison studies. However, if consistent evidence emerges from a logical sequence of studies, the investigators will gain crucial information about the differences between groups on clinically relevant measures. Natural comparison studies are vital components of the researcher's methodological tools.

Case-Control Designs

Researchers using this design select a group of participants or patients who have a specific disease or disability (case) and compare them with a control group. The cases and the controls are compared on previous exposure to defined risk factors such as pathogens, life events or diseases. The method of data collection is usually *retrospective*. To illustrate the use of case-control designs, let us look at a simple hypothetical example. Say that there is an outbreak of food poisoning in a community with 20 people diagnosed with this condition. These people constitute the 'cases'. The control group would consist of people in the community who are not suffering from food poisoning. The hypothesis driving the investigation is that the outbreak was due to a specific unsafe food which was consumed by the affected individuals.

Data collection would involve interviewing the cases to identify the source of the contamination. Say that 18 of the 20 cases had eaten oysters at 'Oscar's Oyster Bar' in the previous 2 days. It is also found that only two of the healthy controls had eaten at Oscar's, but had eaten food other than oysters. The analysis of this hypothetical data would clearly indicate the risk factor for the food poisoning outbreak. After further analyses of Oscar's Oyster Bar and samples from the cases for pathogens (e.g. salmonella), the hypothesis would be confirmed, and Oscar would have to rethink his seafood supplier.

This simple example understates the complexity of the multiple and interacting determinants of health outcomes, in particular chronic diseases such as stroke or cancer. However, case-control designs are relatively simple to implement and provide useful evidence which requires further verification.

Correlational Designs and Outcome Prediction Studies

The aim of correlational studies in the health sciences is to identify interrelationships among variables. In epidemiology, these studies are sometimes referred to as *ecological* designs. Epidemiologists refer to correlational designs as ecological designs. As correlational designs, they aim to establish associations between groups or variables (e.g. association between socioeconomic class and the incidence of type 2 diabetes). A correlation is a statistic that expresses numerically the strength of association that exists between two or more variables.

Table 8.1

Percentage overweight	Number of participants	Number suffering coronary heart disease	Probability of coronary heart disease
Underweight or normal	600	30	$30/600 = 0.05$
10–19% overweight	200	20	$20/200 = 0.10$
20–29% overweight	100	15	$15/100 = 0.15$
30–39% overweight	75	20	$20/75 = 0.27$
≥ 40% overweight	25	10	$10/25 = 0.40$

Let us look at a simple illustration of correlational or ecological study. Say that clinical observations indicate that people who suffer from coronary heart disease tend to be overweight. Such observations might generate the hypothesis: ‘There is a positive correlation between being overweight and the probability of coronary heart disease’. Here, the investigator will need to draw a representative sample of the population of interest (let’s say 500 men and 500 women randomly selected from a population of healthy men and women aged 40 years living in a specified district). The next step would be to measure the participants’ weights and heights. These measurements might be monitored over a period of time to check for drastic changes in weight. The second variable would be measured by the criterion of whether or not the participant suffered from heart disease during a specified period (e.g. 10 years), representing the length of the study. The incidence of coronary disease can then be converted into a probability for a particular category of weight ([Table 8.1](#)).

It can be seen from [Table 8.1](#) that the higher an individual scores on one variable (percentage overweight), the higher the scores on the other variable (probability of coronary heart disease). In this way, the fictional data presented in [Table 8.1](#) are consistent with the predictions of the hypothesis.

Two points should be considered at this stage. First, no evidence has been presented that one variable is causally related to another. This type of investigation does not, by itself, allow us to conclude that, for instance, ‘being overweight causes coronary heart disease’. There are several alternative hypotheses, such as ‘stress causes both being overweight and heart disease’, which can also account for the findings. Second, it is clear that, at least for the period of the investigation, the variable ‘percentage overweight’ does not account for all of the other variable ‘probability of coronary heart disease’. That is, according to the data presented in [Table 8.1](#), there are normal or underweight individuals who suffer from coronary heart disease and very overweight people who do not. Clearly, there must be other variables which influence the incidence of coronary heart disease. There are many other variables, such as smoking, blood cholesterol level, personality type, stress and family history of heart disease, that also correlate with the probability of coronary heart disease.

Some diseases (e.g. lung cancer and chronic back pain) have complex, interacting causes. Natural comparison, correlational and ecological designs can help us identify risk factors: that is, factors that might elevate the risk of the onset or progress of these illnesses in individuals and populations.

Quasi-Experimental Research Designs

If preventive interventions are to be undertaken to reduce the risk factors associated with a disease, then we require reasonable evidence that these factors are, in fact, causally related to the disease. *Quasi-experimental* and *cohort* designs often are used for this purpose. The term ‘*cohort study*’ is preferentially used in the domains of epidemiology and public health. These designs can resemble experiments, with the important difference that there is no random assignment into treatment groups. However, the investigator can control the time and place in which a treatment is introduced or withdrawn. One such method is a time-series design. This type of design is used by both clinical and epidemiologic researchers. Usually, these designs

are *prospective* in that the participants are assessed at the beginning and then followed and assessed over a period of time.

Time-Series Cohort Designs

Time-series designs involve repeated observations before and after the administration of a treatment or intervention. In this way, changes in the sequence of observations after the introduction of a treatment may represent the effects of the treatment on the observed variable. Let us look at an example illustrating the use of time-series designs in health sciences research.

Returning to the risk factor of ‘being overweight’ we discussed previously, the investigation presented here using a time-series design could provide evidence for a causal relationship between this variable and cardiac disease.

1. Select an appropriate population to study.
2. Specify the dependent variable, that is, some clear-cut measure of ‘coronary heart disease’. A commonly used measure may be the incidence of the disease. By ‘incidence’ we mean the number of new cases of the disease reported in relation to the population within a specified period of time (e.g. 50 per 100 000 per year).
3. Introduce an appropriate treatment that reduces the magnitude of the risk factor. In our example, a health promotion package could be introduced, emphasizing exercise and good eating habits. Let us assume that this intervention is adequately financed and a significant proportion of the population adheres to the program. It could then be hypothesized that introduction of the health promotion package will result in a decrease in the incidence of coronary heart disease in the community.
4. Monitor the dependent variable over a period of time. It is essential to have readings of the variable both before and after the introduction of treatment. In this instance, the incidence of coronary illness would be determined from the medical records of hospitals, clinics and physicians. Public health authorities often gather and make available such statistics.

Epidemiologists are involved in the monitoring and surveillance of the health of human populations to track the occurrence of existing and possibly new health problems. The existence of HIV/AIDS was established very early by epidemiologists at the US Centers for Disease Control and Prevention using sophisticated monitoring and surveillance systems. At the heart of such systems are sophisticated data-gathering procedures.

Although the language and concepts of epidemiology are different from the classical clinical disciplines, the approaches to data collection are quite similar to those disciplines. Epidemiologists rely upon the collection of (especially longitudinal) survey data as the basis for their analyses of the distribution and transmission of diseases, they are involved in the analysis of RCTs and, like clinicians, attempt to build causal and statistical models of disease occurrence.

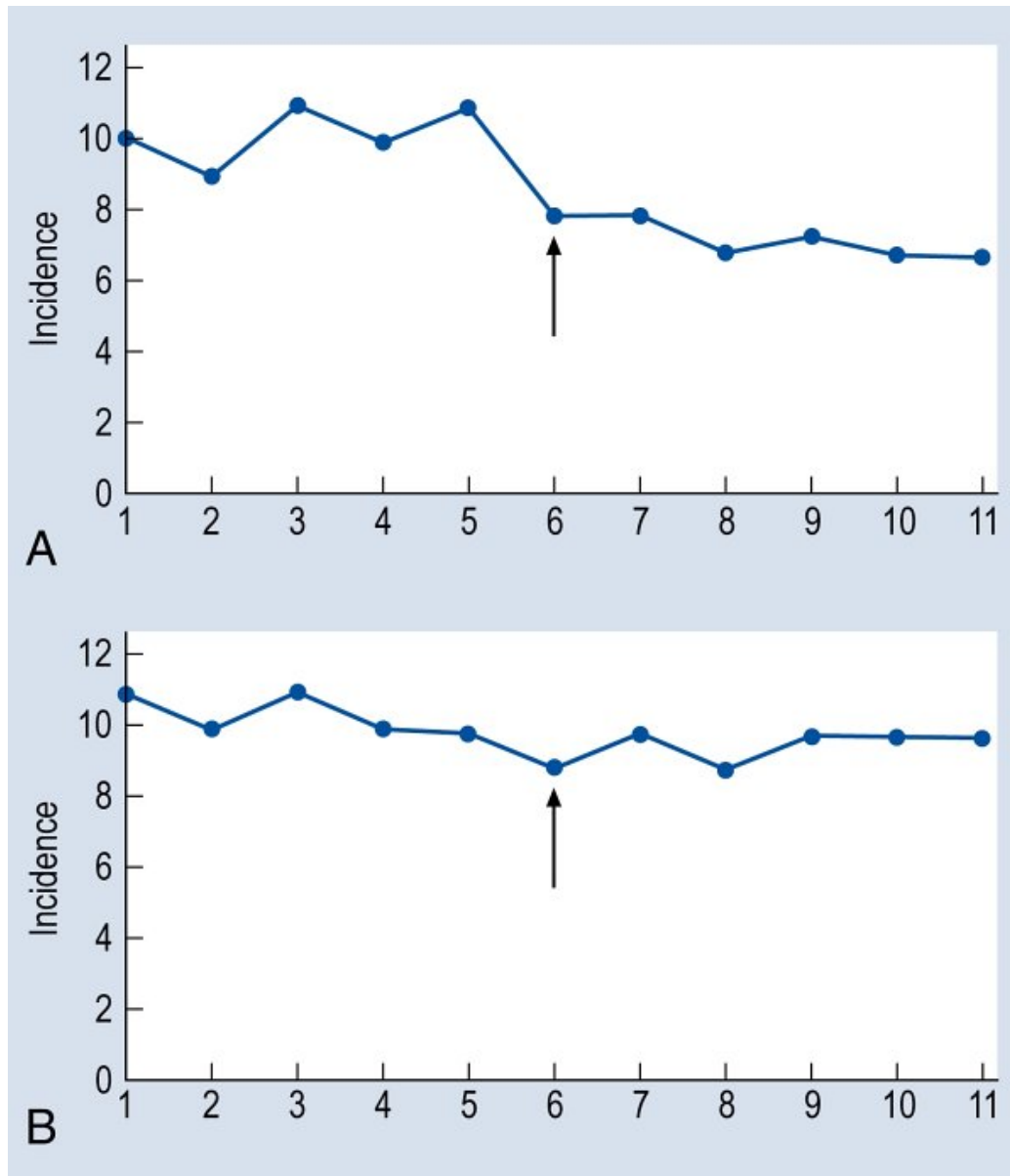


FIG. 8.2 Possible outcomes of time-series designs.

Fig. 8.2 represents two of the many possible empirical outcomes using time-series designs. We will assume that the incidence of the illness was monitored for 6 years before and after the introduction of the treatment.

There are, of course, other possible outcomes. However, in this case, Fig. 8.2A would be consistent with the predictions of our hypothesis, whereas the outcome shown in Fig. 8.2B would be inconsistent with our hypothesis. Discussion of the way in which data generated by time-series designs are analysed is beyond the scope of the present text. Briefly, it involves the

analysis of trends (increase or decrease) found in the dependent or outcome variable.

Time-series designs may have problems with internal validity. What assurance have we that the decrease in the incidence of coronary heart disease, shown in [Fig. 8.2A](#), was caused by the introduction of the health promotion program? Perhaps there was another cause, such as the introduction of drugs to control high blood pressure.

Multiple-Group Time-Series Designs

The introduction of multiple-group time-series designs involves the comparison of two or more naturally occurring groups or cohorts. Let us examine a simple illustration of such designs by a fictional further investigation of the coronary heart disease problem. Using a multiple-group time-series design, the investigator would select a community which is as similar as possible on demographic variables (socioeconomic classes, age, size of the community, etc.) to the community or cohort being studied. In this way we would have:

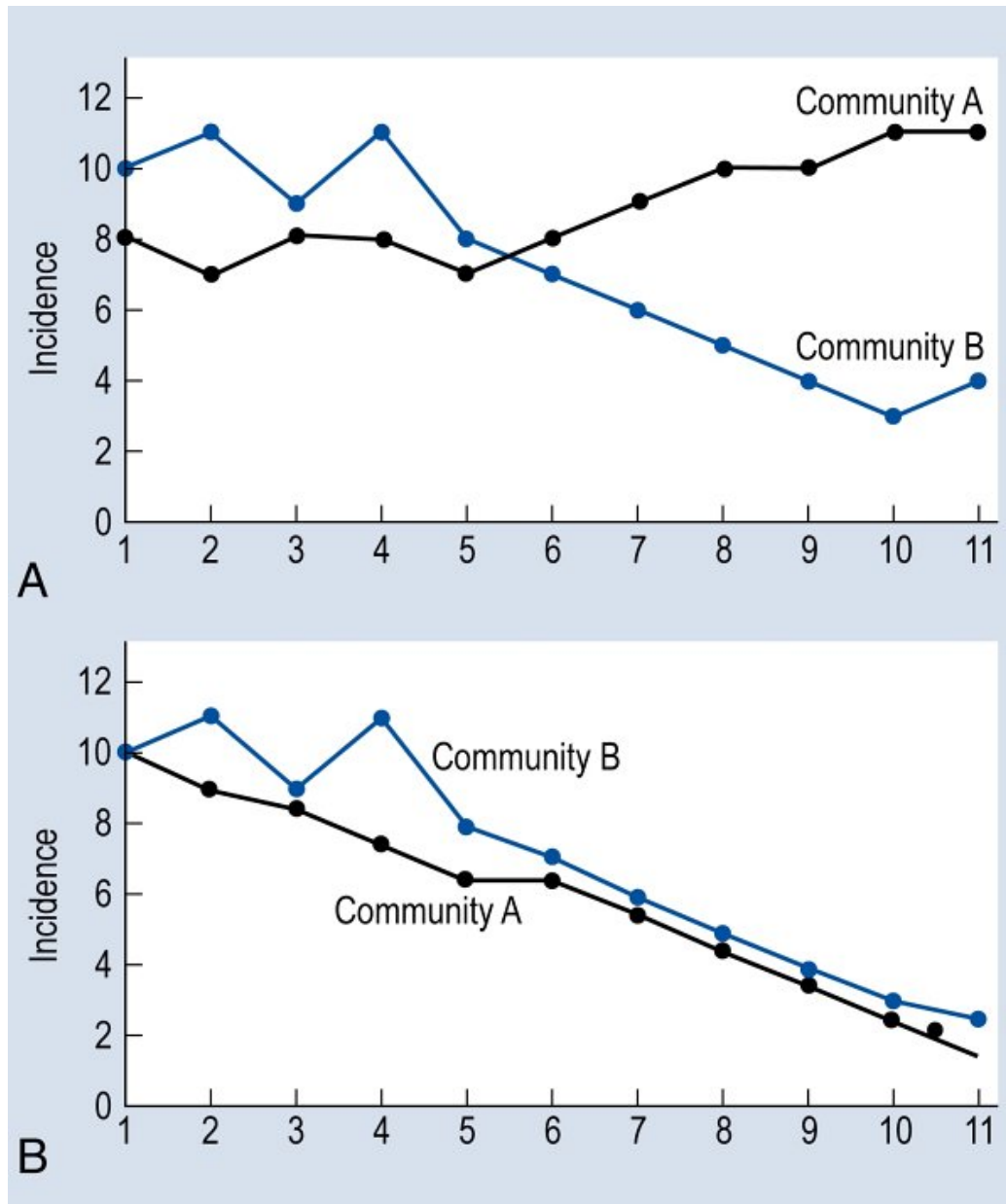


FIG. 8.3 Possible outcomes of multiple-group time-series designs.

- *Community A.* Control. Do not introduce health promotion program.
- *Community B.* Introduce health promotion program.

Fig. 8.3 shows two of the several possible outcomes for such an investigation. Some researchers would call this type of design a prospective cohort design; prospective refers to the fact that the data are collected after

the commencement of the intervention and cohort refers to the fact that we have multiple intervention/non-intervention groups in the study.

Where the two communities show different trends, the outcome shown in [Fig. 8.3A](#) is consistent with our previously stated hypothesis. However, in [Fig. 8.3B](#) there is a trend for decrease in both communities A and B, and therefore the evidence is not consistent with the prediction of the hypothesis.

Even with multiple-group time-series (prospective cohort) designs, some threats to internal validity remain. There are no guarantees that the communities A and B were equivalent on all relevant factors, or that there were no important changes during the study in the communities which might have influenced the incidence of coronary heart disease. The best that researchers can do is to identify and try to estimate the effect of such extraneous factors on the dependent variable. There are problems such as insulating the two groups: when people in community B learn about the program being carried out in community A they, by themselves, might initiate aspects of the program.

The Internal and External Validity of Naturalistic Non-Rct Designs

We have noted several types of problems concerning the internal validity of investigations using natural comparison, correlational and time-series designs. These are similar to the problems of internal validity found in RCT/experimental investigations. However, because researchers have generally less control over the phenomena being investigated, the use of natural comparison designs makes it more difficult to evaluate causal hypotheses. That is, given uncontrolled extraneous variables, a variety of plausible alternative hypotheses might be offered to account for the findings. Therefore in areas such as epidemiology, researchers use evidence arising from a variety of investigations using different types of designs to evaluate their theories and models of the causes of human diseases.

When evaluating evidence from natural comparison designs, the external validity of the findings must be considered. As previously discussed, external validity refers to the generalizability of the results of an

investigation. Strictly speaking, the results of an investigation should be generalized only to the population from which the sample was selected. For instance, consider the example discussed previously, where a sample of male and female students were compared on a test of spatial ability. Any differences between these two groups can be generalized validly only to the population of students from which the samples were drawn. Investigators sometimes forget this obvious point and try to make inferences about males and females in general. Such sweeping generalizations are invalid. For instance, other cultures with alternative childrearing practices might well have males and females with completely different relative spatial abilities. Just because a variable is found to be a risk factor in one community does not guarantee that it will have the same influence on diseases in another community. Clearly, the finding that cigarette smoking is a serious risk factor for coronary heart disease in Western societies does not necessarily mean that cigarette smoking is a risk factor for coronary heart disease among other cultural groups and settings. Given the complex, interacting causes of coronary heart disease, there may be different risk factors in communities which follow different lifestyles or have different physical constitutions. It is not surprising, therefore, that advances in knowledge concerning systematic differences between culturally, racially or sexually different groups have been slow and controversial.

Summary

Descriptive surveys were shown to be important ways of studying the attitudes, behaviours and health problems of a community. Research evidence of this kind can be used to propose theories or test hypotheses of the nature and causes of illness. We examined several types of nonexperimental research designs including epidemiologic approaches, naturalistic comparisons, correlational and ecological designs, and time-series and prospective cohort designs. It was argued that these designs are appropriate alternatives to experimental designs, especially in field settings. However, as with experimental intervention and RCT designs, investigations using these designs must confront the same problems of internal and external validity.

References

1. Canadian Medical Association Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268:2420–2425.
2. Bonita R, Beaglehole R, Kjellström T. *Basic Epidemiology* . 2nd ed. Geneva: World Health Organization; 2006.

Qualitative Research

Introduction,
Research Problems and Questions,
Approaches to Qualitative Research,
What Defines Qualitative Research?,
An Example of Qualitative Research,
Summary,

Introduction

Before we launch into this chapter, it is relevant to briefly consider previous work. We have examined the nature and applications of qualitative methods in earlier chapters. As this work is directly relevant to the present chapter, we have discussed in:

- [Chapter 1](#), the fundamental importance of qualitative method to patient-centred research.
- [Chapter 2](#), the contrast between qualitative and quantitative methods.
- [Chapters 3 and 4](#), the planning necessary for health research in general and the importance of asking the right questions for all methods of enquiry.

- Chapter 5, the sampling strategies relevant for qualitative research.
- Chapter 6, the ethical principles that guide the conduct of all research.
- Chapter 8, data collection.

What was proposed in these chapters is that qualitative (interpretive) research aims to understand people through discovering the meanings of their experiences and actions. The goals of qualitative research are closely related to the goals of patient-centred care. Understanding people requires the study of their beliefs, desires and intentions. ¹ We *infer* what people think and feel by observing and recording what they say and do and take into account the cultural settings in which these actions occur. Qualitative research is well suited for exploring people's experiences of health and disease and their views on the benefits of the treatment and care being offered. The collected data and the evidence obtained is presented in a narrative (i.e. story-like form).

The aims of this chapter will enable you to:

1. Identify the research questions addressed by qualitative research.
2. Describe key approaches to conceptualizing qualitative research.
3. Describe the shared characteristics of qualitative research designs.
4. Explain how qualitative research designs are implemented for producing evidence for advancing medical and health care research.

Discussion of the contemporary philosophy of the health and medical sciences can include a large amount of imprecisely defined terms. However, there is frequent discussion of quantitative and qualitative methods and their merits and appropriate usage.

In a hypothetico-deductive scientific study, the researcher, on the basis of derivations from a theory, *deduces* certain hypotheses. These hypotheses are then tested by comparing predicted results from the hypotheses against actual results. Much of scientific method in the medical and health sciences is based on this approach. It is highly structured and systematic and frequently uses 'objective' measurement tools. A qualitative research study involves an *inductive* approach whereby a range of systematically planned intensive observations are systematically analysed. Qualitative research frequently uses observation and the analysis of participant-generated text or

conversation. The analysis then gives rise to a theory to explain them. A considerable amount of scientific activity involves systematic inductive and deductive enquiry in theory generation and validation.

We find the argument that a chasm exists between hypothetico-deductive and grounded inductive theory to be unconvincing. We consider that both approaches involve the generation and validation of theories that are linked to observations. Major points of difference are the starting point (i.e. theory first or observations first) and the direction of the research activity to connect theory and observations. Neither approach is a-theoretical nor is it unsystematic. We consider that the portrayal of qualitative and quantitative studies as antipathetic competitors is not accurate. Indeed, many studies use the two approaches in a complementary way within the one study. Such studies are called multi- or mixed-methods studies. Examples of such studies are included here.

Research Problems and Questions

When there are significant differences in the cultural backgrounds and experiences of individuals, the understanding of personal meanings becomes problematic. There are numerous areas of health care and medicine where the interpretation of personal meanings is essential to ensure the implementation of effective practices. The four examples presented here illustrate areas where qualitative research can make strong contributions for clarifying personal meanings.

1. *Evaluating interventions.* Previously (see [Ch. 7](#)), we discussed how randomized controlled trials are used to evaluate the safety and efficacy of interventions. In addition to objective evidence, we should also ask if the intervention resulted in meaningful benefits from the participants' perspectives. In [Chapter 10](#), we will discuss the issue of combining researcher and participant perspectives for determining the efficacy of treatment.
2. *Understanding cultural differences between health workers and clients.* In many countries we live in multicultural societies. There is persuasive evidence that the ways in which people experience their bodies, or events such as childbirth, pain or illness, depend to

a large extent on their cultural backgrounds. It is a problem when health practitioners misconceive their clients' views concerning their illness or injury, possibly resulting in erroneous diagnoses and useless interventions. This is a fundamental tenet of patient-centred care, as previously discussed. A particularly important aim of qualitative researchers is to clarify personal meanings of clients and therapists with regard to health care issues to improve communications and enhance treatment outcomes.

3. *Evaluating the effects of health care environments.* Health care institutions, such as general and mental hospitals, are seen as 'subcultures' that strongly influence the lives of both staff and clients. Individuals with chronic illnesses and disabilities requiring long-term care might come to view themselves and their life situations from an 'institutionalized' perspective. Research findings using qualitative methods in this area have been applied to devise strategies to empower people, such as those with intellectual disabilities, to live and participate in the community.
4. *Relating to people with cognitive, neurologic or psychiatric problems.* Individuals diagnosed with problems such as schizophrenia, intellectual disability or brain disorders may, to some extent, experience themselves and the world in ways different from people living without a mental illness. It is in some cases a problem for health professionals that they are not certain how such individuals experience aspects of their world. To ensure that persons with such severe impairments or disabilities are cared for appropriately and with understanding, health professionals must learn to see things from the patients' perspectives. Well-designed and empathetic qualitative research has provided evidence which has helped clarify the personal perspectives of people with severe neurologic and psychiatric problems.

The field of patient experience research is an interesting example of where these approaches intersect. In the past, a major driver for the study of patient experience was a position of principle that patient experience was a fundamental component of health service quality and it therefore should be monitored and used as a key performance indicator. Almost all national health services agencies now require the regular collection of experience.

Such agencies include the US Agency for Healthcare Research and Quality, the Australian Commission on Safety and Quality in Health Care, the UK's National Health Service and the national health services of many other countries. Almost all the studies include quantitative and qualitative data collection approaches. Standard structured response survey questions sit alongside questions where the participants are encouraged to provide their observations, opinions, recommendations and interpretations.

A detailed discussion of this field is outside the scope of this text, but, briefly, it is very interesting to examine the relationship between what is usually asked in the structured response questions and the responses spontaneously offered in the other components of these studies. Campbell, Roland and Buetow's seminal paper in this field argued that patients' principal concerns were effectiveness of clinical care and effectiveness of interpersonal care and access (whether they can get the care they need).² It is interesting that many qualitative responses from participants are strongly critical of technical effectiveness but many patient experience tools do not invite patients to reflect upon this issue in their structured response questions. Thus the selection of the structured response items can direct respondents away from negatively reflecting upon the utility of their treatments. The content emphasis is upon interpersonal interaction and care scheduling.

It is important to understand that patient experience studies are not simply a nice thing to do to improve one's credentials in patient centredness. A recent major review conducted by Doyle, Lennox and Bell (p.01) concluded that:

*'This study, summarising evidence from 55 studies, indicates consistent positive associations between patient experience, patient safety and clinical effectiveness for a wide range of disease areas, settings, outcome measures and study designs'.*³

When one considers the mandatory requirements in most clinical systems for the conduct of patient experience studies and links that with the clear finding that patient experience strongly influences clinical outcomes, it is clear why we have included these discussions in this new version of our text. Although one can mount an ideological argument for competency in

experience studies, there is a stronger utilitarian argument associated with the influence of patient experience on clinical outcomes and the need for appropriate methodological skills to study the use of this link and to capitalize on it.

The examples provided here are cases where qualitative research is appropriate for clarifying personal meanings and enhancing understanding and communication in health care settings. We should note that personal meanings are relevant to a broad variety of problems not only in the obvious areas just discussed. We will examine these questions relevant to evidence-based medicine and health care in [Chapter 10](#).

Approaches to Qualitative Research

There are several schools of thought that have contributed to the development of qualitative research (e.g. Denzin and Lincoln). ⁴ These different traditions influence the way in which research projects for advancing medicine and health care are designed and analysed. That is, plans for conducting qualitative research can differ with the particular conceptual approach adopted by different groups of researchers. We will outline, very briefly, four of the major approaches to qualitative research.

1. *Phenomenology*. Phenomenology, which is both a system of philosophy and an approach to the psychological and the social sciences, emphasizes the direct study of personal experience and the understanding of the nature of human consciousness. Research in this area involves ‘bracketing’ or putting aside the usual preconceptions and prejudices that influence everyday perception so that we can uncover the pure constituents of conscious experience. Within this framework, conscious experience is seen as the basis for personal meaning as we reflect on our experiences in the context of our goals and purposes. An important concept adopted from phenomenology is the notion of ‘multiple realities’; that is, different people may experience the world in quite diverse ways. This suggests that, to understand the meanings of a person’s actions, we must become adept, through empathy, at seeing things from their point of view.

2. *Symbolic interactionism*. Symbolic interactionists emphasize that a social situation has meaning only in the way people define and interpret what is happening. That is, people do not simply respond to 'objective' aspects of their environments, but rather their actions are guided by their personal interpretations of the situation. It follows that different people, on the basis of their past experiences and their particular social positions, may come to interpret a specific situation in quite divergent ways, and may act in conflicting fashions. The central issue is how symbols, in particular those of language, express meaning and enable communication and understanding between people. These are very important issues in health care, for example, when we address issues in sexual orientation and child care.
3. *Ethnomethodology*. Ethnomethodologists study the processes associated with the way in which people perceive, describe and explain the world. They argue that the meanings of specific actions and events are not necessarily obvious but are in fact rather ambiguous and problematic. People select and apply specific rules and principles to define and give meaning to situations in which they find themselves and to justify their actions in a given situation. Ethnomethodologists assert that we take an enormous amount of cultural context, such as norms and rules, for granted in our everyday communications and social interactions, and we tend to 'bracket' these as obvious or common sense. It must be remembered, however, that when the cultural backgrounds of individuals diverge, the understanding of personal meanings in delivery health becomes less obvious or common sense.
4. *Grounded theory*. This approach was developed by US sociologists Glaser and Strauss in the 1960s. Strauss and Corbin define a grounded theory as one which is '... discovered, developed and provisionally verified through systematic data collection and analysis of data pertaining to that phenomenon'. ^{5(p23)} The core idea is to begin the research process without holding an explicit theory or hypothesis which is to be tested through the data. Rather, the theory emerges gradually as more data are collected. The

underlying logic is referred to as inductive logical approach (see [Ch. 1](#)). This method enables health researchers to develop an understanding of how participants actually experience health and illness, rather than ‘filtering’ these experiences through preconceived theoretical ideas.

Each of the four traditions discussed earlier leads to somewhat different approaches to the conduct of qualitative research. There are many other approaches, based on the philosophic theoretical positioning of the researchers. These interpretive approaches include postmodern, feminist, race theories and other approaches which aim to transform society (see Creswell and Poth).⁶ Qualitative researchers are expected to make explicit which particular tradition was selected to guide their project. At the same time, there are characteristics which are common to quantitative research, as outlined designs.

What defines qualitative research?

Given the different approaches, the question is: What is common among the methods of health research projects which have relied on the various interpretive methods? We have already discussed these commonalities in [Chapter 2](#). Qualitative health researchers believe that society exists as a result of meaningful social interaction and that the social actors themselves define it best. Therefore qualitative researchers work towards empathy with the research participants through the sensitive collection of data and maintaining that commitment to validity by interpreting their perspectives in a manner which fits with their intended meanings. Qualitative research is a disciplined enquiry examining the personal meanings of individuals’ experiences and actions in the context of their social environments. *Qualitative* refers to the nature of the data or evidence collected. Qualitative data consist of detailed descriptions based on language or pictures recorded by the investigator. By *disciplined* we mean that the enquiry is guided by explicit methodological principles for defining problems, collecting and analysing the evidence, and formulating and evaluating theories.

Personal meaning refers to the way in which individuals subjectively perceive and explain their experiences, actions and social environments.

Qualitative research provides systematic evidence for gaining insights into other people's view of the world: 'putting ourselves into someone else's shoes'.

The purpose of a research design is to formulate a plan to enable the collection of the data relevant to answering the research question. Regardless, differences in methodological details, designs for qualitative research projects share certain characteristics.

1. Research questions are about people's experiences, beliefs and the meaning of events from their point of view. Qualitative designs therefore aim to ensure close, personal contact between the researchers and the participants. Where possible, qualitative researchers try to collect data in a natural setting; that is, where the participants experience the problems being researched. Of course, this is not always realistic but at least we try to gather as much information as possible and relevant regarding the context.
2. Qualitative researchers use inclusive language. For example, where quantitative researchers will refer to 'subjects', qualitative researchers prefer to view the individual as a 'participant' or a 'consumer' of health services. However, keep in mind that quantitative researchers study variables and mechanisms, and referring to 'subjects' does not imply that they treat research participants in an uncaring dehumanized fashion.
3. Qualitative designs are associated with purposeful sampling, as discussed in [Chapter 5](#). Designs guide the researcher to look for participants who will provide data to improve practices for achieving optimal outcomes from the perspectives of both the participants and the researchers.
4. Implicit in all qualitative research is that the data represents the authentic communications or self-expressions of the participants. Designs must ensure that participants are approached in a way which inspires confidence in the importance of the project and the trustworthiness of the researcher. ¹
5. The research plan must also ensure that persons conducting qualitative data collection (see [Ch. 13](#)) are skilled in techniques of

effective interpersonal communications. All research traditions (both qualitative and quantitative) take it as given that human beings respond actively to their environments so that the way in which participants disclose personal meanings and values are sensitive to the settings and interpersonal dynamics. Therefore it is essential that during data collection the researcher ‘brackets’ their preconceptions and values regarding the phenomenon being studied.

6. A well-formulated design should ensure that the research process is rigorous, conforming to rules of evidence.⁷ An important consideration is ‘bias’ where the researchers adopt a preconceived theoretical framework for collecting and interpreting data. Bias has different connotations in qualitative research in comparison to quantitative research. As Dawes et al state:

The whole point of the study is usually to look specifically at the experience of the participant and this process will be influenced by the researcher’s perspective. What is key here is that this perspective is clearly recorded.^{7(p139)}

In other words, we must make explicit our political and personal biases as researchers as these may colour the evidence and the subsequent conclusions. There are a number of strategies referred to as ‘participatory action research’ where there is an explicit philosophic commitment to a particular cause such as feminism or gay rights. For example, a researcher who has experienced being a refugee may participate in ‘action research’ aiming to explore the health problems and ultimately to improve the social circumstances of refugees. The point here is that we need to be aware of who we are and that the purpose of our research is to produce accurate and honestly reported evidence. Of course, you may exercise the right to be a political activist, but that is not the same thing as scientific health research.

An Example of Qualitative Research

Here we will present an outline of a qualitative paper that illustrates the role of design in the context of the overall research project. We will follow the

stages for conducting applied health research as stated in [Chapters 2 and 4](#). The paper, authored by Van Lith et al, is titled ‘The lived experience of art making as a companion to the mental health recovery process’. ⁸ This research is an illustration of how qualitative designs are implemented when evaluating therapeutic interventions.

1. *Research problems and questions*. The ‘introduction’ provides the background to the research in the form of a brief review of studies about the relationship between ‘art making’ and the process of recovery in ‘consumers in mental health psychosocial rehabilitation services’. ^{8(p652)} Recovery is seen as a process of personal growth, a journey, and ‘which one develops purpose beyond the symptoms, disability and stigma of mental illness’. ^{8(p652)} Art making is defined as a variety of creative activities provided to consumers, in particular painting, sculpting and drawing. The research problem is implied by the aims of the study (i.e. the lack of specific information regarding the ways in which consumers benefit from art making). The research question addresses the consumers’ experiences of art making and *their* interpretations of the ways in which these experiences contributed to their recovery.
2. *Planning*. The researchers collaborated with ‘two of the biggest psychosocial rehabilitation services in Victoria’. ^{8(p653)} Both of these organizations provide extensive services for people recovering from mental health problems. The participants were initially approached by program managers and then selected by the researchers on the bases of their involvement in art-based therapies. Eighteen people volunteered to participate in the study. As the authors indicated, they were using a ‘purposive sampling strategy’, ^{8(p654)} as this approach best identified the participants most capable of discussing their experiences of art-based therapies.
3. *Ethical permission* was granted by the various organizations, including the university to which the researchers were associated. Consumers also were offered the opportunity to ask questions, assured of anonymity and had the right to withdraw from the study at any stage.

4. *Design.* A qualitative design functions to position researchers in a face-to-face, personal relationship with the participants. An interpretive, phenomenological approach was selected to collect and analyse data. This approach enabled the researchers to understand the ‘lived experiences of individual consumers during their recovery’. [8\(p653\)](#)
5. *Data collection.* In-depth interviews were conducted in the setting of the art studio which the participants found comfortable. The researchers reported that each interview, which was tape recorded, was between 45- and 90-min duration. Participants could bring any of their artwork for facilitating the discussion. According to the researchers, ‘each interview aimed to encourage an emergence of the participants’ narrative about art-making and its relationship to mental health recovery’. [8\(p654\)](#)
6. *Organization and presentation of data.* The tapes of each interview were transcribed to produce a written record of the discussions between the researchers and the 18 participants. The interpretive phenomenological approach used by the researchers does *not* require using statistical procedures essential for making sense of numerical data characteristic of quantitative research.
7. *Data analysis.* The data analysis involved reading each interview several times, and by tabulating key phrases and concepts a number of themes emerged. The authors identified ‘11 overarching themes’ [8\(p654\)](#) which represented the experiences of the participants. These themes were linked as closely as possible to the participants’ actual words. Furthermore, it was found that the 11 themes could be grouped into three ‘meta-themes’ which revealed how the art-making process contributed to mental health recovery.

The researchers concluded that art therapy had many benefits for the consumers, including ‘insight, self-esteem, and confidence; empowerment; social connectedness and social engagement; context; and moving on’. [8\(p657\)](#) When evaluating interventions, qualitative researchers focus on the benefits from the perspective of the participants rather than using the objective outcome criteria of quantitative methods. In [Chapter 22](#), we will take a further look at analysing qualitative evidence.

Summary

Qualitative research strategies include data collection that is aimed at understanding individuals in their social environments. Rather than generating numerical data supporting or refuting clear-cut hypotheses, interpretive research aims to produce accurate descriptions based on face-to-face knowledge of individuals and social groups in their natural settings. The role of the researcher in this context is crucial and usually involves physical and social closeness between the participant and the researcher (see [Ch. 2](#)). In this chapter, we looked at a number of traditions for conducting qualitative research, including phenomenology, symbolic interactionism, ethnomethodology and grounded theory. An important point discussed was that regardless of the tradition followed, there are strong commonalities in the designs of qualitative research projects. Also, it was shown by the discussion of a published qualitative study that the structure and problem-solving approach of health research, as outlined in [Chapter 3](#), is followed in empirical studies. We will examine how to conduct qualitative data analysis in [Chapter 21](#).

References

1. Minichiello V, Aroni R, Hays T. *In-Depth Interviewing* . 3rd ed. Sydney: Pearson Education Australia; 2008.
2. Campbell S.M, Roland M.O, Buetow S.A. Defining quality of care. *Soc Sci Med* . 2000;51(11):1611–1625.
3. Doyle C, Lennox L, Bell D.A. Systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* . 2013;3(1):e001570. doi: [10.1136/bmjopen-2012-001570](https://doi.org/10.1136/bmjopen-2012-001570).
4. Denzin N.K, Lincoln Y.S. *The Sage Handbook of Qualitative Research* . 5th ed. Thousand Oaks, CA: Sage; 2011.
5. Strauss A, Corbin J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded*

- Theory* . 2nd ed. Newbury Park, CA: Sage; 1990.
6. Creswell J.W, Poth C.N. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* . 4th ed. Thousand Oaks, CA: Sage; 2018.
 7. Dawes M, Davies P, Gray A, Mant J, Seers K, Snowball R. *Evidence-Based Practice: A Primer for Health Care Professionals*. 2nd ed. Edinburgh: Churchill Livingstone; 2005.
 8. Van Lith T, Fenner P, Schofield P. The lived experience of art making as a companion to the mental health recovery process. *Disabil Rehabil* . 2011;33(8):652–660.

SECTION 4

Data Collection

OUTLINE

Introduction

10. Questionnaires and Survey Design

11. Mixed Methods

12. Interviewing Techniques

13. Measurement and Observation

Introduction

There are numerous methods available for data collection. The appropriate methods are chosen depending on the aims, design and resources of the research project.

Questionnaires are commonly used with survey designs but they also may be used in other studies such as randomized controlled trials. There are different types of questions in questionnaires ranging from highly structured, standardized scales to unstructured open-ended formats. It is better to use standardized questions to measure clinical and demographic concepts. The science of questionnaire and measurement tool construction (psychometrics) has rigorous requirements concerning how measurement tools should be devised and how they should perform. Questionnaire construction needs to follow these principles.

In this section, we note that an interview is, in a sense, a formal conversation between the interviewer and the person being interviewed. As interviews require the presence of an interviewer, this increases the cost and effort needed to obtain the data compared with, say, the completion of a self-administered questionnaire. We note that the technologies available for the presentation of a self-administered questionnaire may include a paper copy or administration via web-based tools and mobile apps. The presence of an interviewer in an interviewer-administered survey may influence the respondent's answers to match the interviewer's expectations. Expert interviewers and well-designed interviews minimize this bias. Qualitative research studies often employ in-depth interviewing techniques which are preferably carried out in the 'natural' settings (homes, hospital, etc.) in which the respondents are living or receiving treatment.

Observational methods also are commonly used strategies for data collection. They may range from highly structured observational protocols, indicating precisely which behaviours or clinical signs should be recorded,

to unstructured records of the experiences of participant observers, as used in qualitative research.

Depending on how the data is recorded and analysed, interviews and observations may be used for both quantitative and qualitative research. However, the use of technical instrumentation to produce valid and reliable numerical data is most appropriate for quantitative research. A variety of standardized measurement instruments is now available for measuring biological and psychological functions.

Whatever data collection strategy is being used, we must ensure that it is reliable (replicable) and valid (accurate). Otherwise the measurement error due to unreliable and invalid data collection strategies may prevent the researchers from achieving their goals.

Questionnaires and Survey Design

Introduction,
Questionnaire Construction,
Questions and Questionnaire Formats,
 Open-Ended and Closed-Response Question
 Formats,
 Likert Scales and Forced-Choice Response
 Formats,
 The Wording and Design of Questions,
The Overall Structure of Questionnaires,
Summary,

Introduction

Questionnaires and interviews have in common the underlying assumptions that various types of information are most accurately known by the patients and research participants themselves and that their experiences and perceptions are worthy of collection. In clinical practice, patient-centred care is now mandated by most clinical agencies globally. The following statement ¹ from the Australian Commission on Safety and Quality in Health Care neatly summarizes the benefits of this approach.

A patient-centred approach makes care safer and of higher quality. It provides demonstrable personal, clinical and organisational benefits. It also satisfies an ethical imperative – involving patients in their own care and in the planning and governance of the health system is the right thing to do. ^{1(pi)}

Given the prominence now attached to patient-centred care, methodological approaches that enable the better collection of information from patients and participants in health and medical research are of prime interest.

Questionnaires are particularly useful for obtaining information from people. The information may be highly varied in content. It can include clinical data such as previous clinical history, current physical and psychological symptoms and attitudes and events experienced by the participant such as degrees of satisfaction with treatments and information about the use of health services and suggestions for improvements to such systems. Of course, questionnaires may also be used by researchers and clinicians to systematize the collection of any information, not just patient information.

Questionnaires are frequently used in surveys. A survey usually involves sampling a group of people and then collecting information from them or about them. Questionnaires are a very convenient method of systematically collecting this information. However, there are other methods that could be used to collect data in a survey (e.g. direct observation such as in a traffic survey). Sometimes people use the terms ‘*survey*’ and ‘*questionnaire*’ interchangeably, but they are actually different things. A *survey* is a study methodology, whereas a *questionnaire* is a data collection tool. These questionnaire tools can be used in many different types of studies, not just surveys.

Questionnaire design skills are very useful in the clinical setting. This chapter focuses on questionnaires and questionnaire design, since they are frequently used to collect data in health sciences research.

The specific aims of this chapter are to:

1. Introduce basic concepts in questionnaire design.
2. Discuss the construction and administration of questionnaires.

Questionnaire Construction

A questionnaire is a set of questions designed with the purpose of seeking specific information from the respondents. Questionnaires are best used with people who are fluent in written and/or spoken communication who can complete the questionnaire themselves (a self-administered questionnaire) or they can be administered to the participants by an interviewer (an interviewer-administered questionnaire). The design of the questionnaire is crucial to its success. The process of design and implementation of a questionnaire is usually termed '*questionnaire construction*'.

Questionnaire construction usually involves the following steps:

1. *The researcher specifies the information that is being sought.* This may involve considerable thinking and discussion. Inspiration for selection of the required information comes from the investigator's research objectives, discussions with others, reading and other sources. At this stage, the document is typically a list of information yet to be translated into specific question form.
2. *Drafting of the questionnaire.* The researcher next takes the list of information he or she wishes to obtain from the respondent and attempts to devise and draft original questions or to select questions from standardized questions used in other major studies. As discussed later in this chapter, the phrasing and design of the questions and the overall design of the questionnaire are important for the validity of the obtained information. If the questionnaire is badly designed, then the responses obtained may not accurately reflect the real data for the respondents. There are many questionnaires that have been devised and tested to measure certain attributes. It is much better to use a standardized and validated questionnaire than to entirely devise your own. This is because questionnaire construction is actually a science (psychometrics) focused on how to devise good quality measurement questionnaires, and standardized tools have been through rigorous testing procedures. Often it is useful to include basic demographic and clinical data in a questionnaire. Again, it is better to use

standardized questions for basics such as age, gender, height, weight etc. rather than inventing one that cannot be directly compared/benchmarked with others. Texts such as Bowling,² Streiner and Norman³ and McDowell⁴ provide comprehensive reviews of such questionnaires, although the health and medical research literature also has literally thousands of publications about tools to measure specific clinical conditions and symptoms. In addition to condition-specific tools it is now common to include general standardized measures such as Burden of Disease (BoD) measures (Thomas & Browning).⁵ BoD is usually measured by Disability Adjusted Life Years (DALYs) and/or by the related concept of Quality Adjusted Life Years (QALYs). A DALY is a year of healthy life that is 'lost' because of a specific condition or disease. When the DALYs associated with a condition or disease within a population are summed, this is the burden of that disease (BoD). The burden of disease is the number of healthy years lost in a population compared with full health. The QALY is the person's length of life multiplied by a valuation of their health-related quality of life. QALY measures are frequently used in the economic evaluation of health interventions. BoD and QALYs and DALYs provide a strong measure of the effect and cost of a disease and are therefore highly suited to economic analysis in health research.

3. *Piloting/testing of the questionnaire with a small test study sample.* It is wise to pilot or trial a new questionnaire with a small group of the intended respondents and also with clinical or research colleagues, to improve its clarity and remove any problems before conducting the main survey. The pilot respondents may be asked whether the questions were clear, and the flow of the questionnaire administration and its acceptance by participants can be monitored. This provides opportunity for improvement of the questionnaire before it is used in the main study.
4. *Redrafting/finalization of the questionnaire.* If the pilot phase uncovers problems with the questionnaire, it will need to be redrafted to address these problems. If they are of a major nature, it is usual to repeat the pilot phase. If they are minor, the researcher

may make the necessary changes and then proceed to administration of the revised final questionnaire to the full sample of respondents who have been selected to participate in the study.

5. *Administration of the questionnaire to the full study sample.* After the questionnaire has been developed, it is administered to the full study sample of respondents. The responses are then analysed in terms of the researcher's aims and objectives. The questionnaire can be self-administered or interviewer administered by an interviewer or research assistant. In terms of self-administration, paper-based administration is rapidly being replaced by Internet and mobile app systems. There are many survey design and administration tools such as SurveyMonkey, Surveygizmo and platforms such as WeChat that make for efficient collection of survey questionnaire data, although there is still an active discussion about concerns for the validity of Internet administered questionnaires (van Gelder, Bretveld & Roeleveld). ⁶

As with all research, the ethics of conducting surveys and designing questionnaires must be considered. For example, respondents should not be misled concerning the true aims of a survey. A blatant example of unethical conduct is if one is asked to respond to a general 'market survey' and then subsequently finds a high-pressure salesperson on the doorstep or contacts the participants via email or phone. If the survey is said to be anonymous, then it is questionable practice by the investigator to code the forms secretly so as to identify respondents. The follow-up of nonresponders can cause a dilemma; people choosing not to participate in a survey should not be pestered. However, forms are sometimes mislaid or forgotten and it is necessary to follow these up to ensure that a representative sample is obtained.

In clinical research, the ethical issues relating to the possible effect of the contents of the questionnaire on the respondent must be considered. As an example, one of this book's authors was involved in a survey aimed at establishing levels of knowledge of Huntington's disease and certain attitudes of people at risk for the condition. Before the survey was undertaken, a pilot study was carried out to establish whether or not the questions were upsetting to the participants. The actual participants were

randomly selected from a ‘pedigree chart’. However, the questionnaires could not be sent out before it was clearly established that each of the prospective participants already knew that they were at risk of developing the condition. It would have been highly unacceptable and appalling if people, for the first time, learned from receiving this questionnaire that they were at risk of a severe genetic disorder.

Questions and Questionnaire Formats

Questions and questionnaires come in a variety of formats. The researcher must decide which format is the most appropriate for the purpose of the study. Let us first consider the issue of the questionnaire format.

In some instances, researchers will not prepare a formal (‘structured’) questionnaire to be filled in by the respondent, but will design a general interview schedule to guide the interviewer who asks the questions of the participant. (Interviews are discussed in more detail in [Chapter 12](#).) There are costs and benefits in both approaches, as shown in [Table 10.1](#).

The interviewer-administered question schedule approach requires expert interviewers to administer the questions, and this is expensive and time-consuming. Further, interviewer bias has been shown to influence responses, as some respondents may modify their responses to fit in with what they perceive to be the opinions of the interviewer. It is important to note that the structure and content of a questionnaire convey a lot of information about the researcher’s research agenda and potentially their hypotheses to the respondent. The respondent generally has little opportunity to influence the agenda in the case of structured questionnaires. A questionnaire is not really a conversation or dialogue but is essentially a monologue from the researcher to which the potential respondent may or may not respond.

However, the self-administered questionnaire approach is relatively cheap, is less susceptible to interviewer bias and can be administered by post or Internet (although it must be said that both unsolicited mail and Internet surveys traditionally have low return rates). Thus the disadvantages of this approach include higher rejection or refusal rates and much less control over how the response forms are filled out. However, in an Internet-administered questionnaire, it can be programmed so as to require

completion of all or key items. Quality control of responses is certainly a problem with some questionnaires returned by mail. Frequently questionnaires that have been returned may need to be rejected because they are only partially completed or they are poorly completed. Anyone involved in self-administered questionnaire analysis will attest to the sometimes-remarkable talents of respondents in returning incomplete questionnaire response forms.

Having decided whether the questionnaire will be self-administered (paper or Internet) or administered by interviewers, the researcher must then decide upon the format for the individual questions. We now turn to a discussion of the main questionnaire design decisions.

TABLE 10.1**Costs and Benefits of Interviews and Questionnaires**

	Costs	Benefit
Interview schedule administered by interviewer	Expensive to administer; requires expert help Responses much more susceptible to interviewer bias	Lower rejection rate More detailed responses can be elicited Greater control over filling out of response form
Self-administered questionnaire	Higher rejection rate Difficult to elicit detailed responses Less control over how response form is filled out	Cheap to administer Less susceptible to interviewer bias Can be administered by mail

Open-Ended and Closed-Response Question Formats

There are two major question formats: the open-ended and the closed-response types. The distinction between the two is best illustrated by example ([Table 10.2](#)).

In [Table 10.2](#) the first question is an open-ended question, whereas the second question is a closed-response question. In an open-ended question, there is no predetermined response schedule into which the respondent must fit a response. In a closed-response question, the respondent is supplied

with a predetermined list of response options. The advantages and disadvantages of both question types are represented in [Table 10.3](#).

Although open-ended questions may elicit more detailed responses from cooperative respondents (or blanks from uncooperative ones), there are some possible disadvantages associated with this type of question. The responses generated by such questions require a very large amount of effort to encode for data analysis and, when they are coded, tend to give rise to categorical scales. Categorical measures are at the low end of the hierarchy for types of measurement scales that may be used in health research. These scales necessitate the use of less powerful statistical methods. Further, some respondents may take a long time to answer this type of question. Or if they do not want to invest the effort, there is a greater chance of refusal to complete some questions. It is easier for the respondent to ‘tick and flick’.

Of course, to a researcher employing a qualitative orientation, these ‘disadvantages’ may not be seen as such. The opportunity to study respondents’ interpretations expressed in their own words might lead a qualitative researcher to advocate extensive use of open-ended questions. However, in this situation it is more likely that interview techniques rather than a written questionnaire would be employed. Questionnaires, particularly of the self-administered variety, are generally used for convenience and speed, not for depth of analysis.

It is important that the lists of options for closed-response questions are carefully designed by the questionnaire designer. It is very easy to bias responses by restricting the range of answers in this type of question.

This brings to mind a short questionnaire that was distributed by an insurance company to one of the authors ([Table 10.4](#)). Two features are remarkable about this question. First, it does not allow for any answers other than the ones listed. Second, the range of available answers is very limited. What we wanted to say was that the insurers were cheap and reliable (i.e. were likely to pay up in the event of a claim). Clearly, the survey designer had a very blinkered agenda and they had not satisfactorily trialled the questions with a nonmarketing audience. Although the designers may well have obtained the answers they wanted, the answers may not have been the ones the respondents wished to give. In the health sciences, there are often large differences in the ways in which health professionals and consumers approach the same problems. In questionnaire design, it is vital

that the researchers do not impose their own conceptualizations of the situations under investigation to the extent that validity is compromised.

TABLE 10.2

Open-ended	Q1. How do you feel about the standard of the treatment you received while you were a patient at this hospital?		
Closed-response	Q2. How would you rate the standard of the treatment you received while you were a patient at this hospital? (circle one number)	Excellent	1
		Good	2
		Moderately good	3
		Fair	4
		Poor	5

TABLE 10.3

Costs and Benefits of Open-Ended and Closed-Response Formats

	Costs	Benefits
Open-ended	Less structured Responses difficult to encode and analyse using powerful statistical methods Greater time taken by respondent to answer Respondent may find writing an essay more difficult than circling a number	More detailed answers elicited
Closed-response	Less 'depth' in answers May frustrate respondents	Tightly structured Responses easily encoded and analysed Less time taken to collect responses

A further example of this danger is provided by the results of a survey conducted at a major teaching children’s hospital in Australia. The survey was designed to study why many parents chose to stay with their children in the hospital, to plan better facilities and services. One of the questions asked of the respondents was ‘Why did your child come to hospital today?’ The investigators deliberately chose an open-ended response format for this question to tap into the parents’ interpretations of the situation in their own words. What a treasure trove of answers was obtained from this general question! The answers provided considerable insight into the issues of importance to the parents, most of which we had not predicted. Thus the ways in which the questions are asked and the answers sought can have a major effect on the value of the information collected.

TABLE 10.4

Why did you choose XYZ Insurance to insure your car?	Newspaper advertisement
	TV advertisement
	Personal recommendation
	Previous insurance with us

Likert Scales and Forced-Choice Response Formats

In attitudinal questions, two possible response formats may be chosen: the traditional five- or seven-point Likert-type format (e.g. strongly agree, agree, neutral, disagree, strongly disagree) or the four-point forced-choice format (e.g. strongly agree, agree, disagree, strongly disagree). These are illustrated by the examples shown in [Table 10.5](#). The first example is a conventional five-point Likert-type scale. The second is a conventional four-point forced-choice response scale. The advantages and disadvantages are summarized in [Table 10.6](#).

The forced-choice format does not allow respondents to give a ‘middle of the road’ or undecided answer. This is to guard against respondents using an acquiescent response mode. Acquiescent response mode refers to the phenomenon that occurs when respondents give middle responses all the time perhaps because of lack of engagement with the task or they may wish

to conceal their true opinions. Extreme response mode occurs when a respondent never selects an intermediate point on the rating scale. For example, if the respondent always chooses ‘strongly agree’ as their answer to all questions then one needs to be a little wary about the accuracy of these answers.

TABLE 10.5

Likert-type	Q1. My medical practitioner always explains the chosen treatment to me (circle one number)	Strongly agree	1
		Agree	2
		Undecided	3
		Disagree	4
		Strongly disagree	5
Forced-choice	Q2. My medical practitioner always explains the chosen treatment to me (circle one number)	Strongly agree	1
		Agree	2
		Disagree	3
		Strongly disagree	4

TABLE 10.6

Advantages and Disadvantages of Response Formats

Response format	Advantages	Disadvantages
Likert-type	Allows middle ‘undecided’ response	Acquiescent response mode
Forced-choice	Respondent forced to give either a positive or a negative response	‘Undecided’ response not allowed

The Wording and Design of Questions

The construction of good questions is an art, and a time-consuming art at that. To obtain valid and reliable responses one needs well-worded questions. There are a number of pitfalls to be avoided. Some of the following are the common mistakes made in constructing questions:

1. *Double-barrelled questions.* This is where two questions are included in one: for example, ‘Do you like maths or science?’ These questions should be separated so that it is perfectly clear to the respondent (and the researcher) which component of the question is being answered.
2. *Ambiguous questions.* It is important to avoid vacuous words and terms that may mean different things to different people. For example, ‘old people’ may mean everyone above 30 years old to a teenager, but everyone above 60 to a 50-year-old.
3. *Level of wording.* It is important to tailor the level of wording of questions to accord with the intended respondents. Jargon is to be avoided, and it should be established in the pilot study that the respondents will understand the concepts. For instance, asking questions about ‘Trisomy 21’ might be inappropriate, whereas ‘Down syndrome’ could be intelligible. Using double negatives should be avoided. In general, questions should be simple and concise.
4. *Biased and leading questions.* The wording of the question should not lead the respondent to feel committed to respond in a certain way. For example, the question ‘How often do you go to church?’ may lead respondents to respond in a way that is not entirely truthful if they in fact never go to a church. A better way to phrase this question would be perhaps to ask about religious worship (e.g. ‘How often do you engage in religious worship?’ or ‘How often do you attend a place of religious worship?’) rather than (Christian) church attendance. Not only can the wording of a question be leading but the response format may also be leading. For example, if a ‘never’ response was excluded from the available answers to the previous question, the respondent would be led to respond in an inaccurate way. Bias might also arise from possible carry-over effects from answering a pattern of questions. For example, a questionnaire on health workers’ attitudes to abortion might include the questions ‘Do you value human life?’, ‘Do you think people should be able to decide about their own medical procedures?’, followed by ‘Do you think abortion should be permitted under any

circumstances?’ In this case the respondent is being led by the context in which the previous questions are asked.

Finally, it should be kept in mind that even a good questionnaire might be invalidly completed. For example, a survey on ‘attitudes to migration’ might be answered less than honestly by respondents who may not want to reveal their negative attitudes to the researcher if the researcher is obviously of immigrant background.

The Overall Structure of Questionnaires

Questionnaires may be structured in different ways, but typically the following components are included.

1. *Introductory statement.* The introductory statement describes the purpose of the questionnaire, the information sought and how it is to be used. It also introduces the researchers and explains whether the information is confidential and/or anonymous.
2. *Demographic questions.* It is usual to collect information about respondents, including details such as age, sex, education history and so on. It is best to position these questions first as they are easily answered and serve as a ‘warm-up’ to what follows. It is also best to use standard questions from other famous studies or from government census questionnaires.
3. *Factual questions.* It is generally easier for respondents to answer direct factual questions, for example ‘Do you have a driver’s licence?’ than to answer opinion or complex and detailed questions. Often, this type of question is also positioned early on in the questionnaire to serve as a warm-up.
4. *Opinion questions.* Questions that require reflection on the part of the respondent are usually positioned after the demographic and factual questions.
5. *Closing statements and return instructions.* The closing statements in a questionnaire usually thank respondents for their participation, invite respondents to take up any issues they feel have not been satisfactorily addressed in the questionnaire, and provide information on how to return the questionnaire.

It is best to avoid complicated jumping and referral structures involving, for example, many conditional questions such as ‘If you answered yes to Question 6 and no to Question 9, please answer Question 10’. Conditional questions usually confuse respondents and ought to be avoided where possible. If the questionnaire is in a self-administered paper format complicated conditional jumps might confuse the reader and they may include or miss material inappropriately.

Summary

Questionnaires are frequently used for data collection in health sciences research. This chapter has reviewed the principles of questionnaire design, including issues arising from the selection of appropriate questions and response formats.

References

1. Australian Commission on Safety and Quality in Health Care. *Patient- Centred Care: Improving Quality and Safety Through Partnerships with Patients and Consumers* . Sydney: ACSQHC; 2011.
2. Bowling A. *Measuring Health: A Review of Subjective Health, Well-Being and Quality of Life Measurement Scales* . 4th ed. London: Open University Press; 2017.
3. Streiner D.L, Norman G.R, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use* . 5th ed. New York: Oxford University Press; 2015.
4. McDowell I. *Measuring Health: A Guide to Rating Scales and Questionnaires* . 3rd ed. New York: Oxford University Press; 2006.
5. Thomas S.A, Browning C.J. Burden of disease. In: Pachana N, ed. *Encyclopedia of Geropsychology* . Berlin: Springer-Verlag; 2016.

6. van Gelder M.M, Bretveld R.W, Roeleveld N. Web-based questionnaires: the future in epidemiology? *Am J Epidemiol* . 2010;172(11):1292–1298. doi: [10.1093/aje/kwq291](https://doi.org/10.1093/aje/kwq291).

Mixed Methods

Introduction,
Perspectives to Health Problems: An Example,
Another Example of a Mixed Method Study,
Sequencing of Qualitative and Quantitative
Designs in Mixed Research Studies,
Examples of Mixed Methods Research
Studies,
Quantitative/Qualitative Designs,
Quantitative/Qualitative Designs: Parallel,
Summary,

Introduction

In [Chapters 7](#), [8](#), and [9](#), we examined research study designs that are suitable for what are sometimes referred to as *monomethod* approaches to data collection. That is, the researchers select either quantitative or qualitative designs for collecting the data relevant to answering the research question. In this chapter, we will outline an alternative to monomethods, referred to as *mixed research designs*. Tashakkori and Teddlie defined mixed designs as:

These are studies that are products of the pragmatist paradigm and that combine the qualitative and quantitative approaches within different phases of the research process. ^{1(p19)}

You may recall that the pragmatist paradigm was discussed in [Chapter 1](#) as a cornerstone for our approach to introducing research methods for health care. Pragmatism in the present context refers to the need to understand health and illness as being the outcome of causal factors that can be quantified and the experiences of people with the health problems. That is, as we discussed in [Chapter 2](#), effective health and medical care is based on evidence from both quantitative and qualitative sources. In studies that are specifically designed to include both methods, we use the term ‘*mixed method designs*’.

Perspectives to Health Problems: an Example

In [Chapter 2](#), we discussed the ‘social perspectives of reality’ using different perspectives when approaching a particular medical issue. Here let us look at a more relevant medical example, a severe and progressive neurologic disorder, Parkinson’s disease (PD). PD is a complex disease characterized by a variety of symptoms, most strikingly motor impairments such as tremors, rigidity and slowness of movement which lead to activity limitations and social participation. Currently, there is a strong effort in the advancement of the understanding and treatment of PD. Iansek and Morris provide a multidisciplinary, patient-centred approach to research and practice to ensure that people with PD receive optimal services. ²

Movement disorders affect both families and individuals. For this reason, contemporary models of service provision are comprehensive and recognize the need for diagnoses, assessments and treatments to be tailored to the individual needs of people living with movement disorders and the significant others in their lives. Comprehensive service provision means enabling people to access skills and advice from a wide range of health and social care disciplines with

coordinated and complementary contributions from different team members. ^{2(pVII)}

Table 11.1

Study phase	Frequency (days gambled in a 4-week period) Mean (sd)	Time (hours spent gambling in a 4-week period) Mean (sd)	Spend (net loss AUD in a 4-week period) Mean (sd)	Gsas total Mean (sd)
Baseline (t_0 , n = 297)	17.82 (16.73)	35.21 (43.26)	\$4,320 (6,457)	26.40 (8.05)
Post treatment (t_1 , n = 260)	9.57 (10.77)	16.41 (25.18)	\$2,365 (11,592)	18.41 (8.90)
6 months (t_2 , n = 249)	10.29 (11.80)	15.46 (23.50)	\$2,000 (5,507)	18.38 (9.37)
12 months (t_3 , n = 235)	9.82 (11.43)	16.49 (23.35)	\$1,891 (7,943)	17.89 (10.24)
Delta 12 months- baseline	-8 (-45%)	-18.72 (-53%)	-\$2429 (-56%)	-8.51 (-32%)

Mixed designs are particularly relevant for research conducted in the context of multidisciplinary patient-centred approaches.

The complex pathology of diseases, such as PD, requires a biopsychosocial approach to research. There are three fundamental considerations when selecting research designs for studying complex disorders:

1. The specific aspect of the condition being investigated (i.e. biology/neuropathology or psychology/activity limitation or social participation, quality of life).
2. The professional position of the researcher (e.g. neuroscientist, neurologist, nurse or physiotherapist).
3. The specific research question being asked.

For example, a neuroscientist working in a laboratory may investigate the question of whether a newly discovered drug reduces the accumulation of α -synuclein, a process which leads to the gradual loss of cells in PD. This researcher or research team would rely on experimental, quantitative designs to monitor the level of α -synuclein accumulation in the brains of animal models of PD. This would be a monomethod approach.

On the other hand, a physiotherapist may introduce a modified exercise program to improve the activities of daily living in patients with PD. The

benefits of the program could be evaluated using both qualitative and quantitative mixed methods to evaluate efficacy from the perspectives of both the professionals (researchers) and the patients (participants).

Another Example of a Mixed Method Study

The PROGRESS (The PROblem Gambling RESearch Study) trial was a large (n=297) mixed methods design, with a parallel group, pragmatic randomized controlled trial as the primary component, and embedded qualitative studies conducted alongside.³ The purpose of PROGRESS was to conduct a superiority study of four different treatments for problem gambling and to determine the durability of the clinical effects. Primary outcomes for this study were changes in gambling symptom severity and gambling behaviours (frequency, time and expenditure). Secondary outcomes were changes in depression, anxiety, stress and alcohol use. Standardized tools were used to assess all of these variables. [Table 11.1](#) summarizes the study's quantitative results.

Two additional qualitative data collections were undertaken with a large stratified random sample of the main study population. One was focused on treatment experiences and the other on the small dropout group that did not complete the full course of treatment and baseline, post-treatment, 6-month post-treatment and 12-month post-treatment data collections.

The study showed robust and durable treatment effects, and the qualitative and the quantitative data provided very useful complementary accounts of the experiences of the participants. The qualitative analyses reinforced the importance of patient treatment and recovery intentions and self-efficacy/confidence in undertaking the treatment.

Sequencing of Qualitative and Quantitative Designs in Mixed Research Studies

There are different ways to implement mixed methods studies depending on the question being investigated. Tashakkori and Teddlie provide an overview of the combinations of qualitative and quantitative designs which have been commonly applied in health science research ([Fig. 11.1](#)).¹

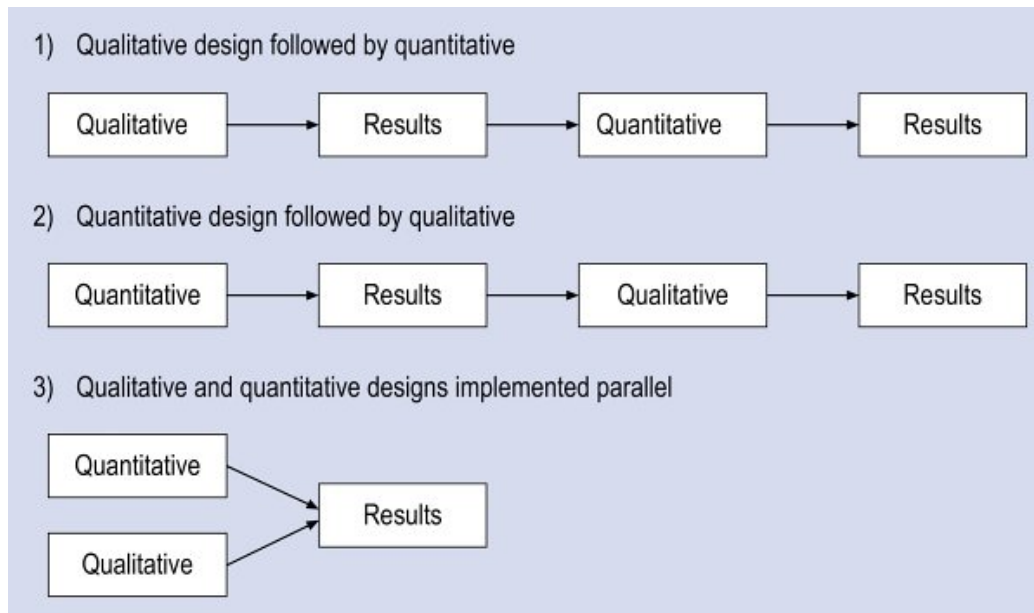


FIG. 11.1 Sequencing of qualitative and quantitative designs in mixed research studies

In the next section, we will look at examples of sequencing the qualitative and quantitative designs.

Examples of Mixed Methods Research Studies

One of the authors has conducted a study of how people evaluate primary health services.⁴ The first step in this process was to conduct focus group interviews with 20 groups of eight participants specifically selected from a wide range of ethnic backgrounds, ages and sexes. The groups were conducted by a facilitator who presented questions concerned with knowledge and opinions of and satisfaction with health services. The discussions were recorded and transcribed.

One set of analyses of the transcripts involved consideration of everything that had been said about the health services with regard to satisfaction or dissatisfaction. This resulted in a range of separate categories or themes. These themes, therefore, were directly derived from the participants' own words and interpretations of their experiences. The themes were then framed in the form of questions that sought information from people about their satisfaction and dissatisfaction with health services.

The questions were then incorporated into a questionnaire. When the questionnaire was piloted with a sample of 500 patients who attended several doctors' surgeries over a period of 3 weeks, it was found that none of the participants nominated new factors that affected their satisfaction and dissatisfaction. Thus the procedure used in developing the questionnaire had very effectively captured how people decided whether they were satisfied or dissatisfied with their health service. This study is an example of where qualitative and quantitative research methodologies can combine powerfully.

Quantitative/Qualitative Designs

It is a well-known problem that some parents do not fully immunize their children and thereby expose them to the risk for potentially life-threatening diseases such as measles and meningitis. A useful research question is why is it?

First, we may conduct a quantitative survey to provide objective information regarding the magnitude of the problem in a defined community. That is, we will obtain clear statistics regarding the proportion of parents who choose not to vaccinate their children even when this community service is freely available.

Second, we may conduct qualitative research to try to understand the reasons why parents decide not to vaccinate children. Berry and colleagues conducted a qualitative survey of general practitioners and community nurses working in New South Wales, Australia, to explore their experiences with 'vaccine-objecting' parents.⁵ The qualitative, in-depth interviews with the immunization providers were the sources of narrative data which identified strategies to best implement immunization for their children with parents who refuse this intervention.

Other qualitative research may directly interview parents who refuse to vaccinate their children so that we can understand their point of view. Ultimately, qualitative and quantitative research will give health providers the evidence which will inform the best practices for protecting children against preventable communicable diseases.

Quantitative/Qualitative Designs: Parallel

The previous two examples involved mixed designs where the evidence was collected sequentially. In the first example, subjective information was provided by focus group interviews. Subsequently, the themes which emerged from the qualitative data were employed to construct a quantitative questionnaire to evaluate people's satisfaction with the provision of primary health services. In the second, the hypothetical quantitative survey for identifying the prevalence of parents refusing vaccinating children was followed by a qualitative study to explore the experiences of providers to deal with parents who refuse vaccination for their children.

Here we will examine a study by Fisher and colleagues in which both qualitative and quantitative methods were used to evaluate the acceptability of body-worn sensors for monitoring some of the motor symptoms of PD.⁶ These sensors are rather like apps you might have on your mobile phone to monitor your steps but are more sophisticated and record the components of movement disorders. The research problem addressed was that there was no rigorous evidence showing whether wearing these sensors was acceptable to the patients.

The researchers used three different methods for collecting data:

1. Quantitative: a structured questionnaire.
2. Quantitative: recording estimated 'no wear' time.
3. Qualitative: a 'free text' response on the questionnaire to report experiences in own words.

The results from the three data sources were analysed to evaluate the acceptability of using the sensors. *Triangulation* refers to comparing and integrating the evidence from different sources. In this study, the combined results clearly demonstrated that patients found wearing the sensors acceptable.

The three studies discussed here demonstrate that qualitative and quantitative methodologies can be combined to provide a more integrated view of the evidence for solving a health problem. This is particularly the case when we are working to understand a complex disease processes, which is associated with disability and problems of living. We need to

understand how the mechanisms of the problem work, as well as the experiences and actions of the patients.

Summary

There are many productive ways for combining qualitative and quantitative approaches to planning and interpreting research projects. As we have argued, care effectiveness studies and the planning are profitably informed by both quantitative and qualitative approaches. Interested readers should consult Tashakkori and Teddlie.¹

We have considered a number of research designs in the previous four chapters. These designs are standard, proven plans for collecting data when conducting quantitative and qualitative research. There is no implication that one design is intrinsically superior to another. Rather, it is the nature of the research problem being investigated and the questions we are asking which determine the optimal design. Also, as discussed in previous chapters, the resources available to complete the project and the ethical treatment of research participants also shape the design of a project. The ultimate criterion for the relevant research design is that it enables access to the best possible evidence to solve health problems and to implement safe and effective practices.

We mentioned Ianssek and Morris as an important overview of current multidisciplinary practice in the area of PD.² James Parkinson, a physician living in London, first described PD in 1817. Over the past 200 years, it is safe to say that every research design which we have described in the previous four chapters has been employed to advance the understanding and effective treatment of this disease.

References

1. Tashakkori A.M, Teddlie C.B. *Handbook of Mixed Methods in Social and Behavioral Research* . Thousand Oaks, CA: Sage; 2003.
2. Ianssek R, Morris M.E, eds. *Rehabilitation in Movement Disorders* . Cambridge: Cambridge University

Press; 2013 doi: [10.1017/CBO9781139012942](https://doi.org/10.1017/CBO9781139012942).

3. Thomas S.A, Merkouris S.S, Browning C.J, et al. The Problem Gambling RESearch Study (PROGRESS) research protocol: a pragmatic randomised controlled trial of psychological interventions for problem gambling. *BMJ Open* . 2015;5:e009385. doi: [10.1136/bmjopen-2015-009385](https://doi.org/10.1136/bmjopen-2015-009385).
4. Thomas S, Steven I, Browning C, et al. Patient knowledge, opinions, satisfaction and choices in primary health care provision: a progress report. In: Dossel D.P, ed. *The General Practice Evaluation Program: The 1992 Work-in-Progress Conference* . Canberra: Australian Government Publishing Service; 1993.
5. Berry N.J, Henry A, Danching M, Trevena I.J, Willaby H. W, Leask J, et al. When parents won't vaccinate their children: a qualitative investigation of Australian primary care providers' experiences. *BMC Pediatrics* . 2017 17, 19.
6. Fisher J.M, Hammerla N.Y, Rochester L, Andras P, Walker R.W. Body–Worn Sensors in Parkinson's Disease: Evaluating Their Acceptability to Patients. *Telemedicine journal and e-health: the official journal of the American Telemedicine Association* . 2016;22(1):63–69. doi: [10.1089/tmj.2015.0026](https://doi.org/10.1089/tmj.2015.0026).

Interviewing Techniques

OUTLINE

Introduction,
Interviewing Models,
Methods of Conducting Interviews,
The Interview Process,
Methods of Recording Interview Information,
 Advantages and Disadvantages of Interview-
 Recording Methods,
 Focus Groups,
The Analysis of Interview Data,
 Quantitative Analysis of Interview Transcripts,
 Qualitative Analysis of Interview Transcripts,
 Coding and Thematic Analysis,
Summary,

Introduction

Interviewing is a very commonly used method of data collection in the health sciences. An interview is a dialogue or conversation between the interviewer and the research participant with the purpose of eliciting

information from the participant. Interviews are a key tool for the clinician and the health researcher when collecting information about the opinions and experiences of patients, research participants and factual information. Interviews vary substantially in their structure, content and the way in which they are conducted. This chapter is concerned with the design and conduct of interviews and the analysis of interview data.

The aims of this chapter are to:

1. Distinguish between structured and unstructured interviews.
2. Outline commonly used strategies for conducting interviews.
3. Compare and contrast quantitative and qualitative strategies for conducting interviews.

Interviewing Models

Many researchers distinguish between structured and unstructured interviews. Sometimes the terms '*formal*' and '*informal*' or '*guided*' and '*open-ended*' are also used (e.g. Morse & Field). ¹

Denzin and Lincoln distinguished between three forms of interviews: (1) the schedule standardized interview in which the wording and order of all questions are exactly the same for every respondent; (2) the non-schedule standardized interview where certain types of information are desired from all respondents but the particular phrasing of questions and their order are redefined; and (3) the non-standardized interview in which no prescribed set of questions is employed. ²

If one defines an interview as a conversation, then a schedule standardized interview is a very rigid form of conversation, almost like a play with a fixed script. In its most structured form, a structured interview may involve someone reading a prepared questionnaire to respondents and then filling in an answer form or response sheet based on their answers. The questions are provided in a systematic, usually fixed order, with minimal or no deviation from the prepared script. In a structured interview, the role of the interviewer is to ask the questions and the role of the respondent is to provide the answers with minimal extraneous information. Conversely, an unstructured interview may involve the interviewer asking no direct

questions, but simply prompting respondents to reflect on their current interests and concerns. Clearly, between these extremes lie a variety of different types of interview strategies and degrees of structure. The three different approaches to conducting interviews can be seen as being on a continuum: structured, semi-structured, and unstructured. Semi-structured and unstructured interviews may include in-depth approaches designed to elicit more detailed and personal information from participants (Minichiello et al). ³ The extent of 'structure' or 'formality' is determined by a number of factors, including:

1. *Whether there is a fixed set of questions or schedule.* In a structured interview, the interviewer has a preplanned set of questions or schedule. These questions may be presented in a fixed order. In an unstructured interview, there may be particular 'themes' to be explored without a specific order required or without specific question wordings.
2. *The way in which the information is recorded.* There are a number of ways in which interview information may be recorded. Structured interviews tend to employ preplanned answer sheets or response schedules. Unstructured interviews have fewer expectations and restrictions on the answer formats of the respondents. The interviewer may record the interview using an audio or video recorder or take free-form notes during the interview.
3. *The types of questions.* Structured interviews tend to employ more closed-response questions in which the valid (acceptable to the interviewer) answers have been preplanned, rather than open-ended questions where the respondent determines what is valid to them. With open-ended questions, the respondents provide their answers in their own words, whereas closed-response questions involve a choice of answers provided by the interviewer.
4. *The extent of control by the interviewer.* In a structured interview, the interviewer explicitly guides or directs the conversation (e.g. 'Mr. Smith, let's discuss how your family feels about your problem') rather than the respondent setting the agenda. In an

unstructured interview, the respondent may assume a more active role in the conversation.

Quantitative researchers often favour structured, standardized interviews, whereas qualitative researchers often prefer semi-structured or unstructured approaches. This is consistent with the way data is analysed and interpreted under the two methods (see [Ch. 2](#)). It is useful to consider some of the advantages and disadvantages of the different types of interview approaches. These are summarized in [Table 12.1](#).

The appropriateness of the different interview approaches is determined by the objectives of the researcher. If the researcher simply wishes to collect some basic symptom data, an unstructured interview would not be efficient. However, if the researcher wishes to study people's conceptualization and interpretation of their illness, an unstructured interview may be quite suitable. Some clinical interviews, such as history taking, are highly structured, whereas other clinical interviews, such as those involving management of a long-term problem, may be less so.

Table 12.1

Advantages and Disadvantages of Structured and Unstructured Interviews

	Advantages	Disadvantages
Structured interviews	May be less time-consuming The same information is collected for all respondents	Responses may not be recorded in the respondents' own words
Unstructured interviews	Responses may be recorded in the 'own words' of the respondents, hence less bias through interpretation The respondent has some input into the research agenda	May be time-consuming Not all the same information is collected for all respondents

Methods of Conducting Interviews

An interview is a conversation, thus there are several possible ways of conducting it. The interview may be conducted in person ('face-to-face') or by remote means such as by telephone or by using a system such as Skype. There are a number of advantages associated with face-to-face interviews. Face-to-face interviews permit the nonverbal reactions of the respondent to be observed and perhaps the development of a closer rapport arising from the more 'natural' setting. Interviewers may use their observations of nonverbal cues to supplement the verbal information being provided and use their own nonverbal cues in a similar fashion. This is a particularly important consideration for qualitative researchers. However, the face-to-face interview may require a substantial amount of travel time for

participants and hence may incur higher costs than a telephone interview. With certain interview objectives, however, telephone interviews may not be suitable. If the interviewer and his or her credentials are not well known to the interviewee, it is unlikely that participants in a telephone interview would provide valid and reliable information about personal topics. Some people find disclosure of sensitive information to be easier by telephone or by the use of anonymous self-completion questionnaires. The face-to-face interview may be too confrontational or embarrassing for them. Tools such as Skype provide an interesting combination of the two methods if the video-cam is enabled.

The Interview Process

1. *Defining the participant population to be interviewed.* One of the interviewer's first tasks is to define the type of participants to be interviewed (i.e. specify the participant population). In quantitative research, strategies such as random sampling or stratified random sampling are used to ensure a representative sample. As qualitative researchers, the interviewer selects those who are most likely to provide the required insights into the situation or issue under study (i.e. the 'key informants').
2. *Recruitment of research participants.* The interviewer must enlist the participation of the research participants. Typically, the interviewer will contact the prospective participant, explain the purposes of the interview and make a number of assurances. These assurances may include protection of privacy, the ability to vet materials based on the interview (e.g. reviewing the interview transcript) and specifying the extent of time the research participant will be involved. Sometimes, the interviewer might write to participants first and then contact them via telephone to arrange the interview to be less 'confronting'.
3. *The interview.* The process of the interview varies substantially according to the methodology to be employed by the interviewer. The process of a structured interview is quite different from that of an unstructured interview. However, some basic goals are shared.

The desirability of eliciting the participant's views rather than reflecting those of the interviewer (i.e. maximization of validity) is paramount. To achieve this, interviewers need to be sensitive, nonevaluative, alert and skilled at delivering and sequencing their questions. In in-depth interviewing, multiple interviews may be required. In this type of approach, the emphasis is on depth of analysis with a smaller number of interviewees rather than the wide breadth of coverage of interviewees and larger participant numbers offered by a sample survey. Substantial practice and good interpersonal skills are required to achieve competence in interviewing. Both video and audio recording of interviews have one large advantage over other methods of recording interview information. This is that the interviewer's interpretation of the interviewee's answers is open to independent scrutiny because the primary research materials are available for study by others (with appropriate ethical clearance). This enables minimization of bias that may be introduced by biased recording of responses by the interviewer if there is no actual audio or video recording of the interview.

4. *Use of response schedules/answer checklists.* When conducting an interview, the interviewer may record the information provided by the interviewee on a predesigned response schedule/answer checklist. For example, the schedule may contain information such as the sex and age of respondents and areas for recording their answers to particular questions. Typically, the response sheet is completed during the interview, although it can be completed at some time after the interview. Immediate recording is probably more valid and reliable although, once again, the presence of a recording device may be of concern to the respondent and may stifle their responses. Unless the response sheet is well designed, there is a major problem of flexibly handling novel or unexpected turns in the interview. The interviewer is interpreting, on the fly, the answers and information provided by the interviewee. If those do not conform to the assumptions designed into the response sheet, these interpretations may not be satisfactory or well considered. Telephone interviews may involve the use of computer assistance

where the interviewer is prompted by, and records the responses on, a computerized schedule.

5. *Free-form (unstructured) notes.* This method of recording involves interviewers making free-form notes to record information they believe to be salient, either during or after the interview. This method of recording is used extensively by clinicians in case notes. There are a number of advantages and problems with such recording; these result from the process whereby the interview is distilled into the notes. This process involves substantial judgement and interpretation on the part of the interviewer. Such distillations result in highly refined (or biased) and reduced data, the validity of which, at least in terms of the interview, is inaccessible to scrutiny. Furthermore, free-form note-taking may not result in the recording of the same type of data across interviewees. Often, when case audits from records are being performed, it is not possible to derive the required data from clinical notes because of this problem. For the participant, it can be a bit off-putting, watching the interviewer write down all the answers. So, although a recorder may be initially intrusive, it is far less intrusive than watching the interviewer write down everything that is said during the interview. Some respondents find this to be intimidating and distracting.
6. *Follow-up.* Having completed the interviews, the interviewer may wish to follow up with the participants. Some interviewers undertake to give the respondents copies of their transcribed interview and may offer the right of vetting the materials included in the transcript. However, this also creates dilemmas in that although the notes or transcript may be an entirely valid record of what was said, what if the respondent subsequently develops concerns about the consequences of what they truthfully said? It is an interesting conundrum.

Methods of Recording Interview Information

Interviewers may use a number of different means of recording interview information, ranging from written summary notes of the interview to an

actual video or audio recording of the live interview. These recording methods have a number of advantages and disadvantages, as discussed here.

1. *Video recording.* The cost of video technology has decreased and the method has improved to such an extent that high-quality video recordings of interviews are well within the budgets of many organizations and individuals. Video recordings provide a wealth of information about the interview. It is possible to observe nonverbal communication channels and to construct audio transcripts of the interaction between the participants. However, some interviewees find the presence of the camera to be threatening. This may have a number of effects: one may be to refuse to participate, another may be to alter the normal flow of the interview. Some respondents may be unwilling to commit personal information and/or controversial views to tape, if they believe there is a risk that the interviewer might disclose the interview to others or that the recording may 'fall into the wrong hands'. Interviewees would find it difficult to deny their views when they have been videotaped expressing them. This is why the police in many countries videotape their interviews with suspects. They cannot be denied later.
2. *Audiotaping.* Many interviewers use audio recordings of interviews to be able to prepare transcripts for later study. Most mobile devices (smartphones and pads) now have this capacity although without an external microphone, recording quality may be poor. Many of the same issues that pertain to video recording are also relevant to audio recording. The use of audio recording may result in greater refusal rates and the 'sanitization' of the views expressed by participants for fear of reprisals arising from disclosure of the interview to others. However, participants often consider audio recording to be less intrusive than video recording. Note, however, that it is not ethically acceptable to record an interview in other settings without full disclosure and agreement from the participants. Qualitative research methodologists have developed systematic methods in note-taking and coding of interview

information. For a more detailed discussion of these issues, the reader is referred to Minichiello et al.³

Advantages and Disadvantages of Interview-Recording Methods

The advantages and disadvantages of the various recording methods are summarized in [Table 12.2](#). Thus video recording is intrusive, requires substantial post-interview analysis and may result in less disclosure, yet it provides very rich information that can be independently analysed. The use of a recording method such as a response sheet requires great trust in the judgement and recording abilities of the interviewer. There is potential for bias arising from the interviewer ‘adjusting’ the information provided by the interviewee to fit the recording method and/or expectations of the interviewer. Furthermore, there is no opportunity for reanalysis.

So which information-recording method for interviews is the most appropriate? The appropriateness of the method of recording interview information is determined by the needs of the person using the information. For example, if the information user simply wants some basic data such as the age, sex and symptom profile of a patient, it would be absurd to use video recording. This would be very time-consuming. Each tape would have to be made, and then viewed again for analysis. In this instance, a simple response sheet or checklist would suffice. However, if the interviewer was interested in exploring reactions of interviewees to the death of a close relative, perhaps the use of audio or video recording would provide a richness of data suitable for that interest.

Focus Groups

Focus groups are a qualitative research technique that is ideally suited to eliciting values, preferences and ideas from groups of people. They are therefore ideal for experience studies and quality improvement studies. A focus group is a qualitative research method ‘that involves a discussion among a small group of people including a moderator or facilitator’.^{4(p7)} The facilitator introduces some open-ended questions for discussion and

moderates the discussion to ensure that it stays on track and that participants are respectful and allowed to contribute.⁵ The discussion of the group is recorded, and then transcribed into written verbatim transcripts. The transcripts are then analysed to understand their content. The analysis may be conducted in different ways according to the qualitative analytical models used by the researcher. A focus group is *not* a group interview. It is a dynamic, focused, interactive discussion.

Table 12.2**Advantages and Disadvantages of Different Ways of Recording Interview Information**

	Advantages	Disadvantages
Video recording	Full transcripts of interview possible Nonverbal data available Accessible to independent analysis	Intrusive Less disclosure Necessity for substantial, and costly, post-interview analysis Potentially greater rates of refusal to participate
Audio recording	Full transcripts of interview possible Accessible to independent analysis	Intrusive (but probably less than video) Reduced disclosure Necessity for substantial, and costly, post-interview analysis Potentially greater rates of refusal to participate
Response sheets	Same data recorded for all interviews Little post-interview analysis required, reducing costs	Unexpected answers may not be well handled Interviewers may bias data in their recording Inaccessible to independent analysis

	Advantages	Disadvantages
Unstructured notes	Cheap and simple	Interviewers may bias data in their recording Some data may be omitted Inaccessible to independent analysis Necessity for some post-interview analysis

Robert Merton (1946) originated the focused interview that then developed into the focus group. The use of focus groups is not limited to health research. They have been used extensively in product and service development in business. ⁶

The role of the facilitator is essential to an effective focus group; they must be skilled at controlling the flow or discussion in a manner which enables participants to contribute as much as possible and yet ensure that the direction of the discussion remains within the parameters of the research question. As with any data collection method, focus groups are not without bias and the researcher must remain mindful of any distortions or prejudices arising from group discussions.

A focus group requires a list of open-ended questions to be put to the participants. The questions are not an interview schedule. They are meant to be used by the facilitator to prompt or provoke discussion. Questions that can be answered by a simple ‘yes’ or ‘no’ are poor focus groups questions. Such dichotomous questions will shut down, rather than promote, the discussion.

A well-designed focus group question list is not extensive. This is not an interrogation. We have conducted focus groups with normally 4 to 10 questions. The Hjorleifsson, Hammer and Diaz study had one question only, ‘What strategies do you use when meeting immigrant patients?’ ⁷ A common mistake among researchers new to focus groups is a desire to have

many structured questions. This generally results in a poor-quality focus group.

A focus group requires a skilled facilitator to be effective. Managing a group of participants requires experience and skills.

The facilitator needs to be able to ensure that an opinionated participant does not dominate the discussion. This requires good technique. The facilitator generally has several stock phrases and techniques to do this. For example, they may say to a dominating participant: 'Thank you (participant name), that is very interesting. I wonder what (another participant's name) thinks about this?' To a timid participant they may say: 'We would like to hear everyone's opinion. Do you have any thoughts about this, (participant's name)?' The goal of the facilitator is to ensure that all of the participants contribute to the discussion.

The facilitator must be neutral and not lead the participants. Poor technique will involve facilitators sharing their own personal experiences and opinions with the group. This directs the group and pollutes its purity. An inexperienced facilitator may inadvertently communicate disapproving body language to some participants if they are not skilled in projecting neutral body language. The researcher and the facilitator are not meant to be studying themselves. A focus group is not a conversation between the researcher and the participants. A unique feature of focus groups is that it is not a one (facilitator) to many (participants) relationship. If it is well conducted, it involves interactions between all the participants not just with the facilitator. It is common to have multiple focus groups in a study. There are several reasons for this. It is usual to have 8 to 12 participants in a focus group. Too small a group may result in hesitant and unreflective discussion. But, on the other hand, too large a group may have the same result. It is very difficult to facilitate large groups.

Table 12.3

Authors	Participants	Groups	Study purpose
Laue et al 2016	GPs from seven different countries	7	How GPs implement COPD guidelines
Yung et al 2016	Chinese adult patients	6	Patient experience of insomnia
Li et al 2015	Hong Kong Chinese women patients	15	Smoking behaviour attitudes and experience
Zhao et al 2014	Chinese GPs	3	Job training needs for GPs
Shin et al 2012	Chinese patients and physicians	7	Smoking perspective of incorporation of smoking cessation into tuberculosis care
Schiotz et al 2017	European patients with multimorbidity	3	Quality of care for patients with multimorbidity
Hjorleifsson et al 2017	GPs in Norway	3	GP strategies approaches to immigrant consultations
Tan et al 2017	Older Singaporeans with type 2 diabetes	4	Experience with diabetes self care management
Skeje & Brekke 2015	TCM practitioners in Shanghai	1	Low rates of referral of infants for acupuncture

A notable feature of the use of focus groups is the diversity of applications to which they are put. Table 12.3 lists the characteristics of some recent international studies that have used focus groups. The studies were deliberately eclectically selected to illustrate the wide range of study participants, group numbers and study purposes.

The Focus Group Question Schedule

A focus group requires a list of open-ended questions to be put to the participants. The questions are not an interview schedule. They are meant to be used by the facilitator to prompt or provoke discussion. Questions that can be answered by a simple yes or no are poor focus groups questions. Such dichotomous questions will shut down, rather than promote, the discussion.

A well-designed focus group question list is not extensive. This is not an interrogation. We have conducted focus groups with normally 4 questions to 10 questions. The Hjorleifsson, Hammer and Diaz study listed previously had one question only: ‘What strategies do you use when meeting immigrant patients?’. ⁷ A common mistake among researchers new to focus groups is a desire to have many structured questions. This generally results in a poor quality focus group.

The Analysis of Interview Data

The manner in which interview data may be analysed is determined in part by how the data have been recorded and, in part, by the theoretical

orientation of the researcher. In the previous discussion, we showed that these formats include videotapes, audiotapes, completed response sheets and free-form summary interviewer notes.

The basis for many analyses of interview data is the interview transcript. The transcript of an interview is a verbatim written version of the conversation that took place between the participants. To provide an example, an excerpt from an interview transcript produced by Janet Doyle at La Trobe University follows.⁸ The transcript is of an interview between a clinician and a client, concerning the client's hearing loss.

Clinician: Okay. So what are you noticing with your hearing?

Client: Well, in a crowded area I can't, you know, understand the other people.

Clinician: Right, sound's a bit jumbled up.

Client: And when I am in the next room I can't even hear the phone.

Clinician: Right.

Client: I am not bad, it is like I am not that bad but still at times.

Clinician: Right.

Client: There is a lot of times I can't hear it.

Clinician: If people speak directly to you like this you are fairly good?

Client: Yes I am alright.

Clinician: But in a group have a bit of trouble?

Client: Have trouble.

Clinician: Right, does one ear seem better than the other ear?

Client: Yes, this one seems better than this.

Clinician: Right.

Clinician: How long have you been noticing your difficulty?

Client: Oh, about 12 to 18 months I suppose.

Clinician: Just gradual was it?

Client: Yes.

Clinician: Okay. Do you get any ringing or buzzing in your ears?

Client: Oh, now and again, very seldom though.

Clinician: It doesn't bother you?

Client: No.

Clinician: Okay. Have you had medical trouble with your ears like infection or anything?

Client: No, no.

Clinician: Do you know of any family history of hearing loss?

Client: No, only the older brother, he has got a hearing aid.

Clinician: Right.

Client: That's all.

Clinician: Have you been exposed to excessive noise?

Client: Yes.

Clinician: ...machinery or?

Client: Yeah. I worked down the car plant, you know, with the heavy machinery.

Clinician: The assembly line?

Client: Right.

Clinician: Were you doing that sort of work for a long time?

Client: Oh yes, 30 years.

Clinician: Yeah, that's a fair...

Client: I wasn't on the line all the time but.

Clinician: Right.

And so the interview continued.

Let us now consider some of the analysis options under the quantitative and qualitative headings.

Quantitative Analysis of Interview Transcripts

A number of quantitative analysis possibilities are presented with interview transcripts. For example, the researcher might count the number of words spoken by each participant to obtain a quantitative measure of their relative contribution to the conversational process. Another possibility would be to count the number of questions asked by the clinician. These quantitative measures could then be used to test various hypotheses. Analyses of interview transcripts similar to the example from Doyle and Thomas, in the study from which it was taken, have demonstrated substantial sex differences between the number of questions asked by male and female clients. It seems that the male clients asked many more questions than the females. Thus the quantitative researcher might use interview transcripts to count and analyse certain features of the transcripts.

Qualitative Analysis of Interview Transcripts

Under the qualitative heading there is a broad variety of approaches to the collection and analysis of interview data. Such approaches may, however, be broadly categorized as *descriptive* or *theoretical*.

A descriptive qualitative study is often termed '*ethnography*'. They often are written from the perspective of the participant(s) in the first person. The purpose of the ethnography is to provide a detailed description of a particular set of circumstances and to encourage readers to make their own interpretations. To make such interpretations, it is essential to consider and theorize the social and cultural contexts in which the evidence was produced.

Ethnographic studies are primarily used in health and anthropology and can help researchers understand and address health/social problems from a specific cultural perspective. Some researchers spend decades living in or involved with a particular ethnic community to understand and research what issues are considered important to them.

Most qualitative studies, however, are theoretical in nature. That is, they attempt to develop theories and concepts and, often, to verify these concepts and theories. A key approach to theoretical qualitative research is provided by Glaser and Strauss' grounded theory.⁹ There are two methods for the development of grounded theory: the constant comparative method, in

which the researcher codes and analyses data to develop concepts, and the theoretical sampling method, in which cases are selected purposively to refine the ‘theory’ previously developed. In contrast, however, some qualitative researchers prefer to collect a substantial amount of data before attempting to formulate their theories.

Coding and Thematic Analysis

Coding is used to organize data collected in an interview and, for that matter, in other types of documents such as field notes. Different qualitative researchers advocate different approaches to coding, but it typically involves a number of steps.

The researchers closely study their materials, in this case the interview transcripts, and develop a close familiarity with the material. During this process, all the concepts, themes and ideas are noted to form major categories. For example, in interviews of nursing home residents, some themes that arise might include personal safety, autonomy and decision making, personal hygiene and so on. Often, the researcher will then attach a number or label to each category and record their positions in the transcript.

Coding is an iterative process, with the researcher coding and recoding, as the scheme develops. Some computer programs are now available to assist with the coding analysis of machine-readable transcripts, and these ease some of the clerical burden, although most qualitative researchers still employ manual coding methods.

Having developed the codes and coded the transcripts, the researchers then attempt to interpret their meanings in the context in which they appeared. The reporting of this process typically involves ‘thick’ or detailed description of the categories and their context, with liberal use of examples from the original transcripts. The process of analysis of qualitative research materials is examined in more detail later in this book.

Summary

Interviews may be defined as a conversation between interviewers and respondents with the purpose of eliciting information. Structured interviews generally involve a fixed set of questions or schedule, the use of preplanned

response sheets, a greater proportion of closed-response questions and direction from the interviewer. Unstructured interviews tend not to have these attributes, with less structure and control. Interviews may be conducted face to face or by telephone and both these methods have certain advantages and disadvantages. The focus group is a valuable alternative to the individual interview. Interview data may be recorded in a number of different ways, but the transcript is often used. Transcripts may be analysed using quantitative or qualitative techniques.

References

1. Morse M, Field A. *Qualitative Research Methods for Health Professionals* . Thousand Oaks: Sage Publications; 1995.
2. Denzin N.K, Lincoln Y.S. *The SAGE Handbook of Qualitative Research* . 4th ed. Thousand Oaks, CA: Sage; 2011.
3. Minichiello V, Aroni R, Hays T. *In-Depth Interviewing* . 3rd ed. Sydney: Pearson Education Australia; 2008.
4. Thomas S.A, Steven I, Browning C, et al. Focus groups in health research: a methodological review. *Ann Rev Hlth Soc Sci* . 1992;2(1):7–20.
5. Leung F.-H, Savithiri R. Spotlight on focus groups. *Can Fam Physician* . 2009;55(2):218–219.
6. McQuarrie E.F, McIntyre S. H. focus groups and the development of new products by technologically driven companies: some guidelines. *J Prod Innov Manage* . 1986(3):40–47.
7. Hjörleifsson S, Hammer E, Díaz E. General practitioners' strategies in consultations with immigrants in Norway-practice-based shared reflections among participants in focus groups. *Fam Pract* . 2017;35(2):216–221. doi: [10.1093/fampra/cmz097](https://doi.org/10.1093/fampra/cmz097).

8. Doyle J, Thomas S.A. Capturing policy in hearing-aid decisions by audiologists. In: Conolly T, Arkes H.R, Hammond K.R, eds. *Judgement and Decision Making: An Interdisciplinary Reader* . 2nd ed. Cambridge: Cambridge University Press; 2000:245–258.
9. Liamputtong P. *Qualitative Research Methods*. 3rd ed. Melbourne: Oxford University Press; 2009.

Measurement and Observation

Introduction,
Observation,
Observation and Measurement,
Operational Definitions,
How Do We Operationalize Constructs?,
Measurement Tools and Procedures,
 Reliability,
 Validity,
Standardized Measures and Tests,
What Constitutes a Good Assessment Tool?,
Measurement Scale Types,
 1. Nominal Scales,
 2. Ordinal Scales,
 3. Interval Scales,
 4. Ratio Scales,
Levels of Measurement,
Summary,

Introduction

Observation involves the direct perception and recording of phenomena under study. Many scientific disciplines, such as astronomy, chemistry, anatomy and psychology, began with the careful recording of detailed observations. These observations were (and still are) summarized and represented by accurate pictures and models which remain integral components of our theories and explanations of health and illness. For example, think of models of the circulatory system or diagrams of neural networks which influenced your knowledge of human biosciences. The careful observation and accurate recording of our patients and clients remain integral components of clinical practices. However, contemporary research and practice rely on advanced instrumentation to enhance the reproducibility, detail and accuracy of observation. The development of accurate tools for measurement and observation has been one of the cornerstones of modern science.

Measurement is a key process in health research and in clinical practice. The same issues are important in both settings. If the measurement procedure used in a study or diagnostic procedure is inadequate, its usefulness will be very limited. The same is true in clinical practice.

The development of accurate instruments or tools are the foundations for providing accurate evidence for driving scientific and clinical advances.

The aims of this chapter are to:

1. Explain the relationship between observation and measurement.
2. Discuss key issues in implementing measurement procedures.
3. Describe good measurement procedures in both research settings and clinical practice.
4. Compare and evaluate the use of four different types of measurement scales.

Observation

Contemporary scientific Western medicine has been said to have its beginnings with the discoveries of the anatomist Vesalius and the physician William Harvey. Vesalius was not only a trail-blazing anatomist, but he also

was a renowned teacher. Historically, the careful observation of bodies and their pathologic states generated the evidence which became the building blocks of contemporary medicine and health care.

Observation and Measurement

Observation is not enough to produce the evidence to build medical and health science. We need to transform our observations into precise measurements. The term '*measurement*' refers to the procedure of attributing qualities or quantities to specific characteristics of objects, persons or events. Observations can be quantified in many ways. Consider these examples:

1. *Counting* is an important strategy for creating numerical data. The frequency of occurrence of a particular, clearly defined object or event is recorded and graphed to provide evidence for the incidence and prevalence of health problems, such as autism, or the occurrence of defined symptoms in a patient, such as the frequency of migraine attacks.
2. We can devise simple but useful measurement tools, for example, those used to quantify subjective experiences.
3. As mentioned previously, the development of measurement tools, such as clocks and calendars to measure time or thermometers to accurately measure temperature, have been essential for not only the advancement of science but also for agriculture and economy. The previous decades have seen the development of new measurement tools based on advances in biochemistry, nuclear physics or information science.

Operational Definitions

When researchers begin their projects, they may hold rather vague views of how to assess the theoretical constructs investigated in their study. For instance, if researchers are interested in collecting data on the 'levels of pain' experienced by patients, then they must convert this vague notion of pain into tightly defined statements of how exactly this is to be assessed.

Depending on their theoretical interpretation of the concept of ‘pain’, and the practical requirements of the investigation, one or a combination of the many possible approaches to assessing pain may be selected.

The process of converting theoretical constructs into precise statements of how variables are to be measured is called *operationalization*. It is important that researchers provide exact details of how observations or measurements were taken so that others may judge their adequacy and appropriateness for providing the required evidence. Also, sufficiently clear description of the data collection in process is required to enable others to replicate the procedures in a new study. The protocols followed need to be unambiguously described. To do this may require a considerable amount of detail, so in larger studies it is not unusual to have a separate paper to fully describe the protocol ¹ which sits alongside various results papers. ² A quantitative study that is methodologically adequate in terms of design and sampling may be of limited value if the measurements are inadequate techniques. Let us now discuss operationalization.

The *operational definition* of a construct and associated variables is a statement of how the researcher conducting a particular study intends to measure the relevant variables being investigated. It should be unambiguous and reproducible by other researchers or health professionals.

At the outset, let us note that in most circumstances there is no single, optimal method for measurements. For example, if a researcher claims that her therapeutic techniques significantly increased ‘motor control’ in her sample of patients, the obvious question that arises is, ‘What was meant by “motor control” and how was it measured?’ In our teaching, we often use the example of the eternal evaluation question trilogy. Does it work? What is it? What does ‘work’ mean? If the researcher replied that she was interested in motor control as a score measured on the ‘Plunkett Motor Dexterity Task scores’, she has, in fact, supplied her operational definition. Another researcher may challenge the adequacy of this definition and substitute the assessment suitable for their own research project. However, the selection of operational definitions are not arbitrary, hence researchers invest considerable efforts to develop consistent and applicable definitions.

A useful operational definition will contain enough information to enable another researcher or clinician to replicate the measurement techniques used

in the original study. Furthermore, a good operational definition of a clinically relevant variable will enable a fellow professional to replicate the original diagnostic or assessment procedures.

How do we operationalize constructs?

Accurate and reproducible assessment is contingent on health professionals and researchers establishing clear consensus concerning the ways in which the properties of things are to be identified. If we cannot reach such a consensus then how do we know that we are talking about the same things? The problem is that there are complex and abstract concepts, such as intelligence, attitude, anxiety or pain, that are not directly observable. As discussed previously, the first step is to provide a theoretical framework for defining the construct. The next step is to operationalize the construct or transform the construct into observable and measureable variables, as shown in [Fig. 13.1](#).

Consider, for example, the construct ‘pain’. Most of us have experienced pain at one time or another such that one might think that it is an easily defined experience. However, pain is not something that can be directly observed or measured. Rather, having selected a theory of pain, we need to identify the variables which are the logical manifestation of the construct. The question here is: ‘How do we know that a person is in pain?’ When we hold a theory relevant to describe, we are in effect deducing and specifying the variables which represent levels of pain ([Fig. 13.2](#)).

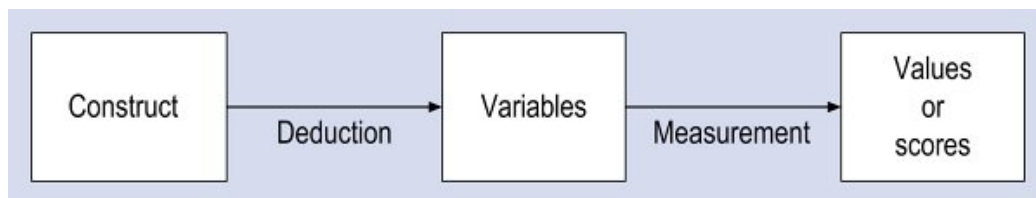


FIG. 13.1 Operationalization of a construct.

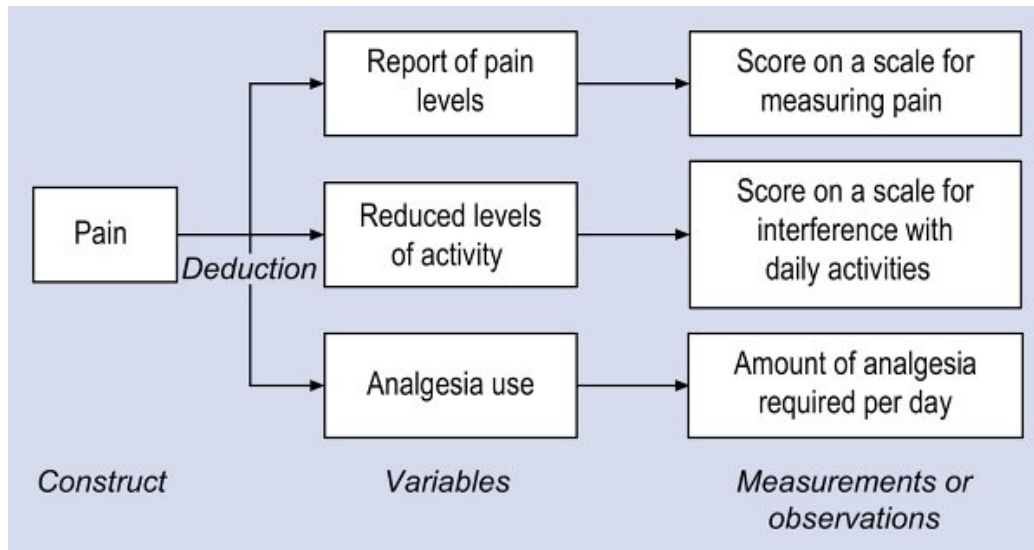


FIG. 13.2 Operationalization of the theoretical construct 'pain'

The *operational definition* of a construct is a precise and unambiguous statement of the procedures (operations) which are necessary for observing/measuring the levels of the variables associated with the construct. Note that a given construct can generate a range of operational definitions (see Fig. 13.2) where we may rely on a number of factors to assess the construct.

For example, Fig. 13.2 shows that the operational definitions of pain are as follows:

1. Subjectively experienced pain levels, reported on a specific scale.
2. Changes in the frequency of behaviours (e.g. going to work).
3. The amount of medication required by client/patient to remain comfortable.

This information is usually found in the 'Methods' section of a research paper, where the details are (or should be) provided concerning the apparatus or instrumentation used and the procedures followed to collect the required information.

Measurement Tools and Procedures

The use of measurement tools is expected to yield scores or values that are reproducible, accurate and applicable to the measurement task in hand. Also, they should be practical; that is, as easy as possible to use. These properties often are given the technical terms of *reliability*, *validity*, *applicability* and *practicability*.

Before these specific test properties are reviewed, it is useful to look at some basic concepts in measurement or test theory. For any measurement outcome, there are three related concepts to assess relevance: the observed value or test score, the true value or test score and measurement error. Hypothetically, if one could be weighed on a completely accurate scale, their true weight would be 100 kg. However, the scales that they use in their bathroom might give them a reading of 90 kg. The difference between the observed score and the true weight, the 10 kg, is the measurement error. This relationship can be expressed in the form of an equation such that:

$$\text{Observed value} = \text{true value} \pm \text{error}$$

Measurement tools are designed with a view to minimize measurement error so that the observed value we obtain from our assessment process is close to the true or real value. There are a number of ways to quantify the reproducibility and accuracy of a data collection procedure.

Reliability

Reliability is the property of reproducibility and consistency of the results of a measurement procedure. There are several different ways in which reliability can be assessed. These include test–retest reliability, inter-observer reliability and internal consistency. Let us examine each.

Test–Retest Reliability

A common way to assess test reliability is to administer the same measurement procedure twice to the same participants. The results obtained from the first test are then correlated with the second test. Reliability is

generally measured by correlation coefficients, which vary from -1 to $+1$ in value. A test–retest reliability of at least $+0.8$ represents a reliability assessment procedure. In a formal study to determine reliability, a sufficiently large sample of subjects would be evaluated two or more times to obtain a correlation coefficient.

Inter-Observer (Inter-Rater) Reliability

A common issue in clinical assessment is the extent to which clinicians agree with each other in their assessments of patients. The extent of agreement is generally determined by having two or more clinicians independently assess the same patients and then comparing the results using correlations. If the degree of agreement (correlation) is high, then there is high inter-observer or inter-rater reliability.

Table 13.1 illustrates examples of both high and low inter-observer reliability on ratings of patients on a 5-point scale. Let's imagine that this scale measures the level of patient dependency and the need for nursing support. As we mentioned earlier, the degree of reliability is quantitatively expressed by correlation coefficients. However, by inspection you can see that in Table 13.1 there is a high degree of disagreement in the two observers' ratings in the 'low reliability' column. In this instance, the clinical ratings would be unreliable and inappropriate to use in the research project. However, the outcome shown in the 'high reliability' column in Table 13.1 shows a high level of agreement.

Internal Consistency

Measurement tools will often consist of multiple items. For example, a test of your knowledge of research methods might include 50 items or questions. Similarly, a checklist designed to measure activities of daily living might have 20 items. The internal consistency of a test is the extent to which the results on the different items correlate with each other. If they tend to be highly correlated with each other, then the test is said to be internally consistent. Internal consistency is also measured by a form of correlation coefficient known as Cronbach's α ; an $\alpha > 0.8$ is considered to be a desirable property for a test.

Table 13.1

Participants	Observers' Ratings			
	Low reliability		High reliability	
	A	B	A	B
1	4	3	4	4
2	2	5	2	3
3	3	4	4	5
4	1	4	2	2
5	3	1	1	1
6	4	4	5	5

Thus the reliability or reproducibility of an assessment or test can be determined in several different ways, including the test–retest, intra-rater, inter-rater and internal consistency methods.

Validity

Validity is concerned with accuracy of the test procedure. Just because one keeps getting the same result with repeated administrations, or agreement among independent observers, doesn't mean that the results are accurate. If one jumps on the bathroom scale and gets a result of 40 kg and then jumps off the scale and then gets back on and it is still reading 40 kg, this reading is certainly reliable, but obviously it is an error (for readers who do not know us, 40 kg as an observed score for our weight entails major measurement error!). Thus the adequate reproducibility or reliability of a test or assessment process is essential, but we also need the results to be accurate or valid.

Types of Test Validity

As with reliability, test validity may be assessed in a number of different ways. These include content validity, sensitivity and specificity, and predictive validity.

Content or face validity

In many contexts, it is difficult to find external measures to correlate with the measure to be validated. For example, an examination in a particular academic subject may be the sole measure of the student's performance

available to determine grades. How can it be determined whether the tests administered will be valid? One way is to write down all the material covered in the subject and then make sure that there is adequate sampling from the overall content of the material delivered on the subject. If this criterion is satisfied, it can be said that the test has content or face validity. You may have had an experience where you felt that a subject assessment task had low content validity because it did not reflect the material presented on the subject.

Table 13.2

Participants	Low Predictive Validity		High Predictive Validity	
	Score on X	Score on Y	Score on X	Score on Y
1	4	3	3	3
2	4	8	4	3
3	5	2	5	5
4	5	7	5	5
5	6	4	6	6
6	8	5	8	7

Predictive validity

Predictive validity is concerned with the ability of a test to predict values of it or other tests in the future. Some tests are designed to assist with prognostic decisions (i.e. what is going to happen in the future); for these tests, high predictive validity is an important quality.

Let us examine an example of predictive validity. Say a researcher devised a screening rating scale, X, for selecting patients to participate in a rehabilitation program. The effectiveness of the rehabilitation program is assessed with rating scale Y. Say that each rating scale involves assigning scores of 1 to 10 to the patients' performance. [Table 13.2](#) illustrates two outcomes: low predictive validity and high predictive validity.

Although the calculation of correlation coefficients is needed to examine quantitatively the predictive validity of test X, it can be seen in [Table 13.2](#) that in the 'low predictive validity' column, the scores on X are not clearly related to the level of scores on Y. On the other hand, in the 'high predictive validity' column, scores on the two variables correspond quite closely. Within the limits of the fact that only six participants were involved in this

hypothetical study, it is clear that only the results in the ‘high predictive validity’ column are consistent with rating scale X being useful for predicting the outcome of the rehabilitation program, as measured on scale Y.

Construct validity

Construct validity is related to the extent to which the underlying theoretical construct is effectively represented by the measurement process. One way to test construct validity is to devise other measures of the construct and to demonstrate that scores on these alternative tests correlate highly with the original test.

Table 13.3

Test result	Real situation	
	Disease present	Disease not present
Disease present	True positive	False positive
Disease not present	False negative	True negative

For example, say that we were to demonstrate that our end-of-semester multiple-choice examination has high content validity in measuring the construct ‘knowledge of research methods’. We could devise another measure, such as carrying out and writing up a brief research project. A high correlation between the scores on the two tasks would indicate a high level of construct validity.

Sensitivity and specificity

The concepts of sensitivity and specificity are most commonly applied to screening tests, where the purpose of the test is to determine whether the patient has a particular problem or illness. There are four possible outcomes for a test result, as shown in [Table 13.3](#).

Sensitivity refers to the proportion of people who test as positive who really have the disease (i.e. the proportion of true positives out of all positives). Specificity refers to the proportion of individuals who test as negative who really do not have the disease (i.e. the proportion of true negatives out of all negative test results). If a diagnostic test has a sensitivity of 1.0 and specificity of 1.0 it is a perfectly accurate or valid test.

Most clinical or screening tests are not perfect and some have unknown or quite low sensitivity and specificity.

A Case Study in Clinical Test Validity

The early detection of breast cancer in women has been recognized as an important public health initiative in many countries. Common ways of detecting suspicious lumps include breast self-examination and mammography, an x-ray of the breasts. Mammography is a common screening procedure and some countries such as Australia have funded large-scale programs to promote it.

However, commendable as these initiatives may be, there are some doubts about the validity of mammography as a diagnostic tool. Walker and Langlands, in an early evaluation, studied the mammography results of 218 women, who, through the use of a diagnostic biopsy, were known to have breast cancer.³ Of the 218 women with cancer, 95 (43.6%) had recorded a (false) negative mammography test result. Of these patients, 47 had delayed further investigation and treatment for almost a year, no doubt relieved and reassured by their 'favourable' test results. The delays in treatment, given what we know about the relationship between early intervention and improved prognosis, in all likelihood seriously compromised the health and ultimate survival of these women. In this instance, the accuracy (or lack of it) of the test results had very important consequences for the people concerned. Test quality is of profound importance in research and clinical practice, as demonstrated by this example. In the past 25 years, research has been completed to improve the accuracy of screening tests. More recently, Evans estimated that improved screening techniques have helped in the early detection of breast cancer and, combined with advances in chemotherapy, have contributed to the survival of thousands of women.⁴ The development of technologically advanced assessment tool is one of the central aims of translational research, as discussed in [Chapter 24](#).

Standardized Measures and Tests

Because reliability and validity of measures are so important, many researchers have devoted considerable time and energy to the development

of measuring instruments and procedures that have known levels of reliability and validity.

The development of measurement standards for physical dimensions such as weight, length and time has been fundamental for the growth of science. That is, we have standards for comparing our measurements of a variable and we can meaningfully communicate our findings to colleagues living anywhere in the world. There are a variety of clinical measures (e.g. the Apgar tests for evaluating the viability of neonates) that represent internationally recognized standards for communicating information about attributes of persons or disease entities.

Furthermore, there are standards relevant to populations, in terms of which assessments of individuals become meaningful. For instance, there are standards for the stages of development of infants: levels of physical, emotional, intellectual and social development occurring as a function of age.

Some tests have been trialled on large samples, and reliability and validity levels recorded. Tests that have been trialled in this way are known as *standardized measures* or *tests*. A large variety is available, particularly in the clinical and social areas. Many US firms and cooperatives market standardized tests for health researchers and practitioners. However, many researchers use tests and measures that have not been standardized, and they do not report levels of reliability in their literature. This is of particular concern in studies where subjective measures with incomplete operational definitions are employed.

What constitutes a good assessment tool?

Considerable effort has been devoted to developing, standardizing and evaluating novel instruments and measurement procedures. Based on our discussion, the criteria listed here are relevant for identifying what constitutes ‘good’ measurement in health research and practice.

- *Theoretical framework.* As there are complex and abstract concepts in the health sciences, researchers need to have a clear theoretical framework in the context of which the measurement procedure is defined and evaluated.

- *The measurement process should lead to reliable outcomes.* This means that the results should be consistent and there should be substantial agreement between researchers or clinicians using a measurement procedure. It follows that if a measuring instrument or a procedure produces unreliable results, then its usefulness is questionable.
- *The measurement procedure should be valid.* This means that we are measuring the targeted, rather than another, construct or variable. Also, the results should be accurate as determined by the criteria for validity discussed in this chapter.
- *The instrument should be accessible.* This implies that the cost and level of expertise required to use the instrument should not be prohibitive. Of course, there are different criteria for accessibility depending on the questions being investigated or the level of resources available to the researcher or practitioner. For example, positron emission tomography (PET) scans are a very expensive and specialized procedure involving the use of radioisotopes. However, the cost and level of expertise required are justifiable in terms of the data produced or the needs of the patients for the accurate diagnoses achieved by PET scans.
- *The measurement procedure should be acceptable to research participants or patients.* It is important to explain, as an ethical requirement, the measurement process and any possible associated risks or burdens (see [Ch. 6](#)). The procedure should be as convenient and safe as possible for research participants.

Desirable Features of Clinical Measures

Easy to administer and interpret, reducing the training expenses and increasing availability of the procedure

The procedure is acceptable to use and safe for patients.

Research evidence has demonstrated the reliability and validity of the measurement procedure.

The procedure can identify most or even all the cases with the condition (i.e. sensitivity).

The procedure does not falsely identify people as suffering from a condition, but instead produces minimal false positives (i.e. selectivity).

Measurement Scale Types

Measurement can produce different types of numbers, in the sense that some numbers are assigned different meanings and implications from others. For instance, when we speak of Ward 1 and Ward 2, we are using numbers in a different sense from when we speak of infant A being 1 month old, and infant B being 2 months old. In the first instance, we used numbers for naming; in the second instance, the numbers indicate quantities. There are four scale types, distinguished by the types of categories, orders or numbers produced by the measurement of a specific variable.

1. Nominal Scales

The ‘lowest’ level of the measurement scale types is the nominal scale, where the measurement of a variable involves the naming or categorization of possible values of the variable. The measurements or recorded observations produced are ‘qualitative’ in the sense that the categories are simply different from each other. If numbers are assigned to the categories they are merely labels and do not represent real quantities; for example, Ward 1 and Ward 2 might be renamed St. Agatha’s Ward and St. Martha’s Ward without conveying any less information. [Table 13.4](#) shows some other examples of nominal scaling.

The only mathematical relationship pertinent to nominal scales is equivalence or nonequivalence (i.e. $A = B$ or $A \neq B$). A specific value of a variable either falls into a specific category, or it does not. Thus there is no logical relationship between the numerical value assigned to its category and its size, quantity or frequency of occurrence. The arbitrary values of a nominal scale can be changed without any loss of information.

Table 13.4**Examples of Nominal Scaling**

Variable	Possible values
Patient's admission number	3085001, 3085002
Sex	Male, female
Religion	Catholic, Protestant, Jewish, Muslim, Hindu
Psychiatric diagnosis	Manic depressive, schizophrenic, neurotic
Blood type	A, B, AB, O
Cause of death	Cardiac failure, neoplasm, trauma

Table 13.5**Examples of Ordinal Scales**

Variable	Possible values
Severity of condition	Mild = 1, moderate = 2, severe = 3, critical = 4
Patients' satisfaction with treatment	Satisfied = 1, undecided = 2, dissatisfied = 3
Age group	Baby, infant, child, adult, geriatric
Cooperativeness with nurse or patients in a ward	(In decreasing order) Mrs Smith, Mr Jones, Ms Krax

2. Ordinal Scales

The next level of observation or measurement involves rank ordering values of a variable. For example, first, second or third in a foot race are values on an ordinal scale. The numbers assigned on an ordinal scale signify order or rank.

With ordinal scales, statements about ranks can be made. Where A and B are values of the variable, we can say $A > B$ or $B > A$. For instance, we can say Mrs Smith is more cooperative than Mr Jones ($A > B$), or Mr Jones is more cooperative than Ms Krax ($B > C$). We cannot, however, make any statements about the relative sizes of these differences. Examples of ordinal scales are shown in [Table 13.5](#).

3. Interval Scales

Interval scales rely on measurements which produce numbers. Examples of interval scales are shown in [Table 13.6](#). For these scales, there is no absolute zero point; rather, an arbitrary zero point is assigned. For instance, 0°C does not represent the point at which there is no heat, but the freezing point of water. An IQ of zero would not mean no intelligence at all but a serious intellectual or perceptual problem in using the materials of the test.

Table 13.6

Examples of Interval Scales

Variable	Possible values
Heat (Celsius or Fahrenheit)	-10°C , $+20^{\circ}\text{C}$, $+5^{\circ}\text{C}$, $+10^{\circ}\text{F}$
Intelligence (IQ)	45, 100, 185

Table 13.7**Variables Measured on Ratio Scales**

Variable	Possible values
Weight	10 kg, 20 kg, 100 kg
Height	50 cm, 150 cm, 200 cm
Blood pressure	110 mm Hg, 120 mm Hg, 160 mm Hg
Heart beat	10 per min, 30 per min, 50 per min
Rate of firing of a neurone	10 per millisec, 20 per millisec, 30 per millisec
Protein per blood volume	2 mg/mL, 5 mg/mL, 10 mg/mL
Vocabulary	100 words, 1000 words, 30 000 words

The use of an interval scale enables identification of equal intervals between any two values of measurements: we can say $A - B = B - C$. For example, if A, B and C are taken as IQ scores, and $A = 150$, $B = 100$ and $C = 50$, then it is true that $A - B = B - C$. However, we cannot say that $A = 3C$ (that A is three times more intelligent than C).

4. Ratio Scales

Ratio scales are characterized by having a nonarbitrary zero point. For example, in the Kelvin temperature scale, 0°K (or absolute zero) represents an absence of heat, in that the molecules have stopped vibrating completely; whereas 0°C is simply the freezing point of water. The centigrade or Celsius zero is an arbitrary one tied to the freezing point of a particular compound. Thus $^{\circ}\text{K}$ is a ratio scale and $^{\circ}\text{C}$ is an interval scale. Examples of variables measured on ratio scales are shown in [Table 13.7](#).

Levels of Measurement

Table 13.8 compares the characteristics of different scales or levels of measurement. Interval and ratio scales represent quantitative measurements. A ratio scale is the ‘highest’ scale of measurement, in the sense that it involves all the characteristics of the other scales and has an absolute zero. A measurement on a higher level can be transformed into one on a lower level in this framework, but not vice versa, because the higher scale measurement contains more information and the values can be put to more use by permitting more mathematical operations than those on a lower level.

Table 13.8

Characteristic	Level of measurement			
	Nominal	Ordinal	Interval	Ratio
Distinctiveness	*	*	*	*
Ordering in magnitude		*	*	*
Equal intervals			*	*
Absolute zero				*

Distinctiveness: different numbers assigned to different values of property.

Ordering in magnitude: larger values represent more of the property.

Equal intervals: same distance between points on a scale.

Absolute zero: zero value represents absence of property.

Also, a given variable might be measured on one of several types of scales, depending on the needs of the investigator. Consider, for instance, the variable ‘height’. This variable could be measured on any of the four scales, as follows:

1. *Ratio scale*. The height of individuals above the ground (e.g. 180 cm).
2. *Interval scale*. The height of individuals above an arbitrary surface (e.g. 100 cm above the surface of a bench).
3. *Ordinal scale*. The comparative heights of individuals (e.g. rank-ordered from tallest to shortest). The scale may be produced by observation or converting interval or ratio scales into rank-ordered categories.
4. *Nominal scale*. Categorizing individuals (e.g. ‘normal’ or ‘abnormal’). These categories can be identified by direct

observation or dichotomizing ‘higher’ (ratio, interval or ordinal scales).

The different types of measurement scales are important when considering statistical analysis of data. Statistics are numbers with special properties obtained by summarizing and analysing data. The type of measurement scale determines the type of statistic that is appropriate for its analysis.

Summary

This section defines good measurement practice and describes its importance in research and clinical practice.

In good measurement practice, we need to define concepts operationally so that other investigators also can carry out or assess the measurement procedure. We also need to establish the reliability and validity of our measurements. A high degree of reliability and validity is necessary for minimizing measurement error. We have noted that measures involving the exercise of human judgement (i.e. ‘subjective’ measures) are not necessarily unreliable or invalid.

Four different scale types were discussed: nominal, ordinal, interval and ratio. These scales have different characteristics, particularly in relation to the permissible mathematical operations. In subsequent sections, we will show that the scale type involved in our measurements determines the descriptive and inferential statistics appropriate for describing and analysing the data.

References

1. Browning C, Chapman A, Cowlshaw S, et al. The Happy Life Club™ study protocol: a cluster randomised controlled trial of a type 2 diabetes health coach intervention. *BMC Public Health* . 2011;11(1):90. doi: [10.1186/1471-2458-11-90](https://doi.org/10.1186/1471-2458-11-90).
2. Browning C, Chapman A, Yang H, et al. Management of type 2 diabetes in China: the Happy Life Club, a pragmatic

cluster randomised controlled trial using health coaches.
BMJ Open . 2016;6(3):e009319. doi: [10.1136/bmjopen-2015-009319](https://doi.org/10.1136/bmjopen-2015-009319).

3. Walker Q.J, Langlands A.O. The misuse of mammography in the management of breast cancer. *Med J Aust* . 1986;145(5):185–187.
4. Evans W.P. Breast cancer screening: successes and challenges. *CA Cancer J Clin* . 2012;62(1):5–9.

SECTION 5

Descriptive Statistics

OUTLINE

Introduction

14. Organization and Presentation of Data

15. Measures of Central Tendency and Variability

16. Standard Scores and Normal Distributions

17. Correlation

Introduction

Data produced in quantitative research is analysed by applying statistics. Depending on the complexity of the research design and the amount of data collected, we will have generated or collected a set of scores called ‘raw’ data. We process the raw data to produce results by using statistical analysis. Section 5 is concerned with descriptive statistics, which represent the mathematical conventions for organizing and summarizing the raw data.

In [Chapter 14](#), we outline techniques used for tabulating and graphing data. These techniques enable us to visualize trends and identify differences across levels of the independent variables. When the data have been appropriately organized we may calculate statistics, such as the percentages and proportions of scores found in the groups being studied.

Another important use for descriptive statistics is to ‘crunch’ or summarize data. The descriptive statistics are discussed in [Chapter 15](#) and include ‘measures of central tendency’ (mode, median and mean) and also measures of variability (range, semi-interquartile range and standard deviation). Using these statistics enables us to summarize the raw data and to efficiently convey information about the research findings for the relevant groups.

Statistics such as the mean and the standard deviation also can be applied to calculating standard scores. Standard scores are used to establish the position of a particular score relative to others in a population. The percentile rank of a score is the percentage of scores in its frequency distribution that are equal to or lower than it. So a clinical test score that is greater than 88% of the scores of the patients taking the clinical test is said to be at the 88th percentile, where 88 is the percentile rank. Percentile ranks are useful in a variety of applications but they are frequently used to understand the ‘meaning’ of a clinical test result. It provides an understanding of whether the score is unusually small, whether it is typical

or average, or whether it is elevated. Whether it is clinically good to have a low, high or average score depends on the particular test but the percentile rank is calculated in the same way for all eventualities. In [Chapter 16](#), we examine how the standardized normal distribution can be used to calculate the position of any specific score within a population and how to interpret the clinical implications of these scores.

In [Chapter 17](#), we examine another important type of statistics called *correlation coefficients*. Correlation coefficients are used to express the degree and direction of association between two or more variables. For example, we could use correlation coefficients to demonstrate if there is an association between the variables ‘level of exercise’ and ‘body weight’. The closer the calculated correlation coefficient is to 1.0 (the maximum value), the more precisely we may predict from one variable to the other. Although correlation coefficients are extremely important for showing how different variables are associated, they do not necessarily indicate causal relationships between the variables. Showing causal effects requires appropriate research designs, as we suggested in [Chapter 7](#).

Organization and Presentation of Data

Introduction,
The Organization and Presentation of Nominal
(Categorical) or Ordinal (Ranked) Data,
 Organization of Discrete Data,
 Graphing of Discrete Data,
Organization and Presentation of Interval or Ratio Data,
 Grouped Frequency Distributions,
 Graphing Frequency Distributions,
Descriptive Statistics for Discontinuous Data,
 Ratios,
 Proportions,
 Percentages,
 Rates,
 Odds,
Summary,

Introduction

The summary and interpretation of data from quantitative research entails the use of *statistics*. A statistic is a number that is obtained by the mathematical manipulation of the raw data. We use descriptive statistics to describe data (e.g. the average and range) and we use inferential statistics to analyse research observations and measurements (e.g. we use inferential statistics to answer the questions ‘are the scores in one group different from the scores in another group?’).

In previous chapters we have examined how interviews, observations and measurement are used to produce data in clinical investigations or research. It is difficult to make sense of raw data when they consist of a list of a large number of measurements or observations. Before we can interpret or communicate the information provided by a research project, the raw data must be organized and presented in a clear and intelligible fashion. We can do this organization process by using descriptive statistics. In this chapter, we will outline methods used in descriptive statistics for the organization, tabulation and graphic presentation of data. We will also examine the use of some simple statistics directly derived from the tabulation of the data.

The aims of this chapter are to:

1. Outline methods for organizing and representing data in the form of frequency distributions, tables or graphs.
2. Demonstrate how the measurement scale used for data collection influences the organization and presentation of the evidence.
3. Discuss the calculation and use of some simple descriptive statistics for discontinuous data, including percentages, ratios and rates.

The Organization and Presentation of Nominal (Categorical) or Ordinal (Ranked) Data

A primary consideration in selecting appropriate statistics is the data type (which may be categorical, ordinal, interval or ratio). Let us examine two hypothetical examples as an illustration.

Organization of Discrete Data

Example 1: Nominal (Categorical) Data

We are interested in the sex of patients (sex is a nominal or categorical data type) undergoing gallbladder surgery (cholecystectomy) at a public hospital over a period of 1 year. The raw data indicating the sex (M or F) of the patients is simply read off the patients' records, as follows:

F, M, M, F, F, F, M, F, M, F, M, F, F, F, F, F, F, F, F, F, M, F, F, F, M,
M, F, M, F, M

Table 14.1

Frequency Distribution of Gender of Patients Undergoing Cholecystectomy at a Hospital over a Period of 1 Year

Sex	<i>F</i>
Male (M)	10
Female (F)	20
	$n = 30$

Grouping the nominal data involves counting the number of cases (or measurements) falling into each category. The total is $M = 10$ and $F = 20$. The data can be presented in tabular form. [Table 14.1](#) shows the following conventions in tabulating data:

1. Tables must be clearly and fully labelled – both the table as a whole and the categories – so that readers can interpret unambiguously what they are reading.
2. f represents frequency (how many) of the cases or measurements fall into a given category.
3. n represents the total number of cases or measurements in a study sample or group.
4. N represents the number of cases in a population. (See [Section III](#) for the difference between samples and populations.)

Example 2: Ordinal (Ranked) Data

Ordinal data are presented by counting the number of cases (frequency) of each ordered rank making up the scale. They also are normally presented in descending or ascending sequence.

An investigator intends to evaluate the effectiveness of a new analgesic versus a placebo treatment. A post-test-only control group design is used: the experimental group receives the analgesic and the control group receives the placebo. Twenty patients are randomly assigned into each of the two groups. Pain intensity is assessed by the patients' pain reports 5 hours after minor surgery on a 5-point scale:

5 = Excruciating pain

4 = Severe pain

3 = Moderate pain

2 = Mild pain

1 = No pain

The raw data are:

Experimental group: 3, 4, 5, 3, 3, 3, 4, 2, 1, 3, 2, 1, 3, 4, 5, 2, 3, 3, 3, 3

Control group: 5, 4, 4, 4, 5, 3, 4, 3, 2, 4, 4, 2, 4, 5, 3, 4, 4, 4, 5, 5

Table 14.2**Reported Pain Intensity of Patients After Placebo and Analgesic Treatments**

Pain Intensity	Experimental Group (Analgesic) F	Control group (Placebo) F
1	2	0
2	3	2
3	10	3
4	3	10
5	2	5
	$n = 20$	$n = 20$

After tallying the results, the data can be presented as a frequency distribution, as shown in [Table 14.2](#). This demonstrates that when the data have been tabulated, we can see the outcome of the investigation. Here, the pain reported by the experimental intervention group is somewhat less than that of the placebo no-treatment control group. Organizing the data is the first step in producing evidence for testing the hypothesis that the new analgesic is more effective than placebo.

Graphing of Discrete Data

Once a frequency distribution of the raw data has been tabulated, a variety of techniques is available for the graphical presentation of a given set of measurements. Frequency distributions of qualitative data are often plotted as bar graphs (also termed ‘*column*’ graphs) or shown pictorially as pie diagrams.

A bar graph involves plotting the frequency of each category and drawing a bar, the height of which represents the frequency of a given category. [Fig. 14.1](#) graphs the data given in [Table 14.1](#).

[Fig. 14.1](#) demonstrates conventions in plotting bar graphs:

1. The y-axis, also called the ordinate, is used to plot frequencies.
2. The x-axis, also called the abscissa, is used to indicate the categories.
3. The bars do not touch each other, reflecting the discontinuity of the measurement categories.

It should be noted that care must be exercised in interpreting graphs, as the axes may be translated or compressed causing a false visual impression of the data. Make sure that you inspect the values along the axes, so that you are not misled. It is also acceptable to calculate the percentage of scores falling into each category and to display the percentages instead of frequencies.

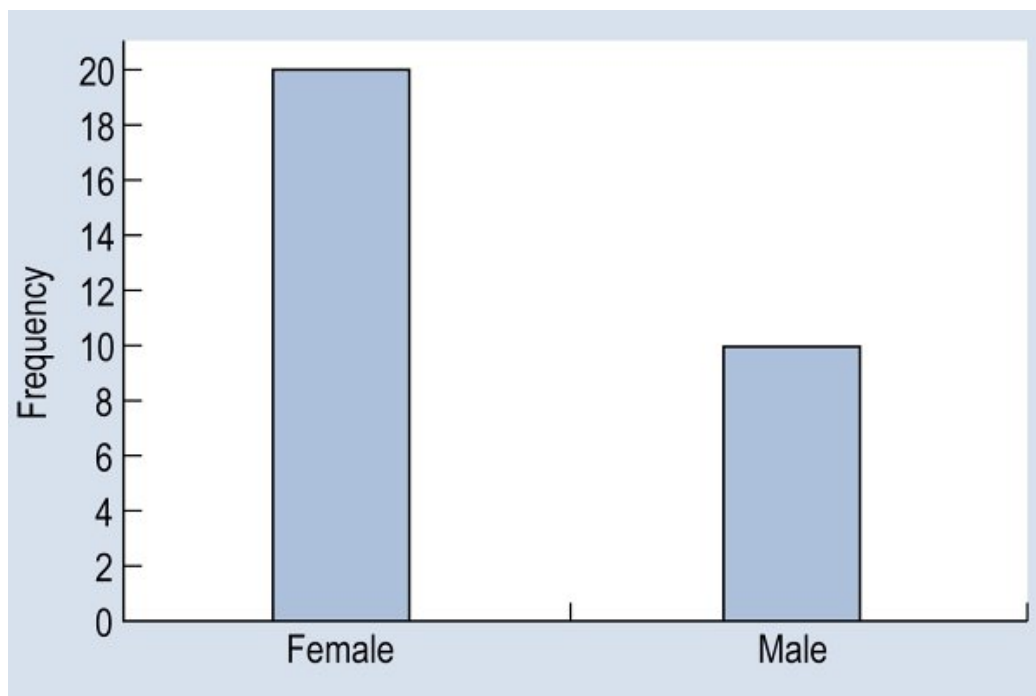


FIG. 14.1 Bar (column) graph of patients undergoing cholecystectomy.

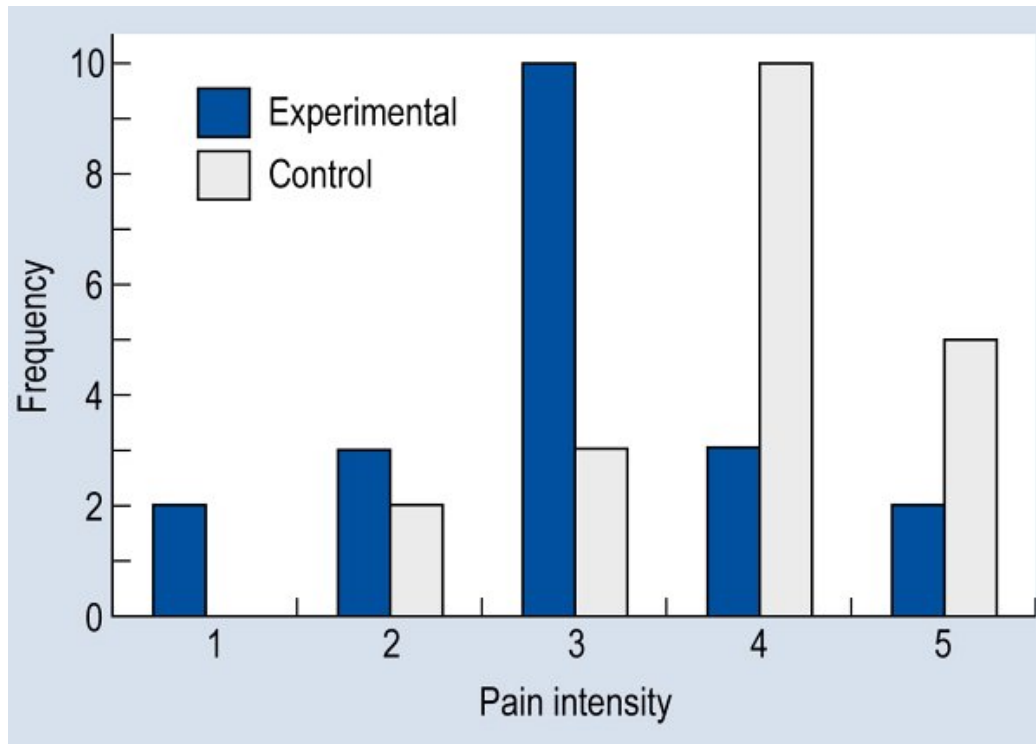


FIG. 14.2 Bar (column) graph of patients' pain intensity.

Fig. 14.2 shows that by presenting the data for the experimental and control groups on the same graph, the reader can gain a visual impression of the possible effectiveness of the analgesic treatment in contrast to that of the control intervention or treatment. Nominal data also can be presented meaningfully as a pie chart, where the percentage of each category is converted into a proportional part of a circle or 'pie'. For example, in a given hospital we have the hypothetical spending patterns shown in Table 14.3.

Fig. 14.3 represents a pie chart of the information. In constructing Fig. 14.3, we converted the numbers into percentages and then into degrees (out of $100\% = 360^\circ$; i.e. each 1% of the total is represented by 3.6° in the circle).

Table 14.3

Hypothetical Spending Patterns

Item	Cost (\$)	Percentage of total
Wages and salaries	1 500 000	50.00
Medical supplies	500 000	16.67
Food and provisions	500 000	16.67
Administrative costs	500 000	16.67
Total	3 000 000	100.01

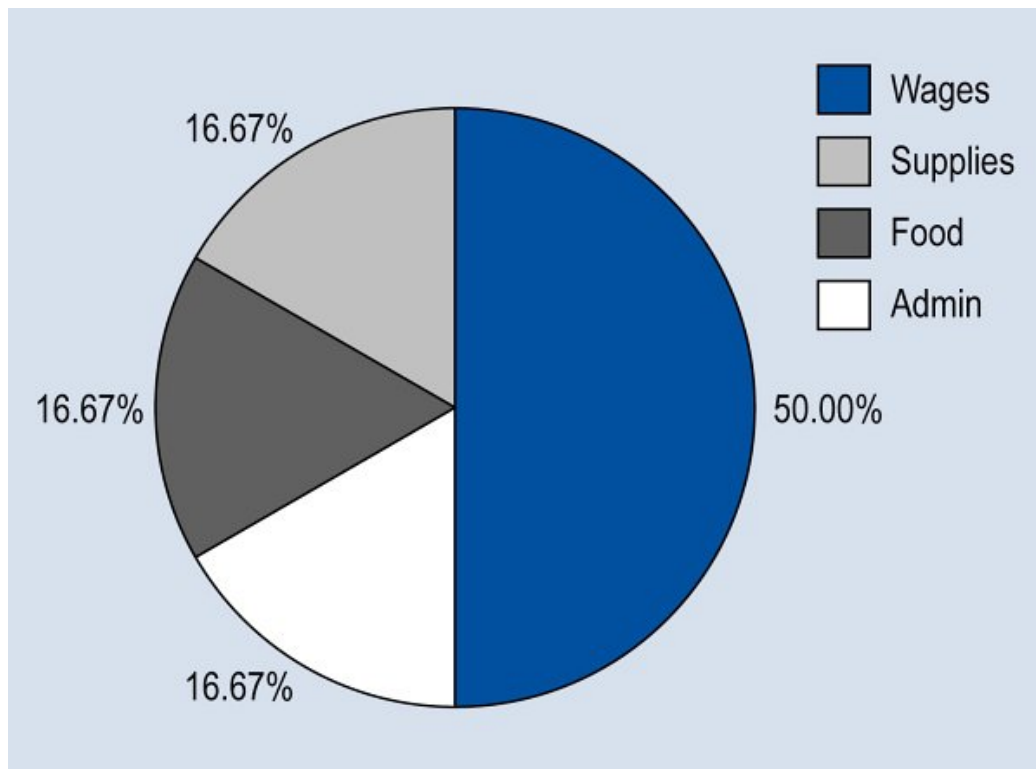


FIG. 14.3 Pie diagram of a hospital's spending pattern.

Organization and Presentation of Interval or Ratio Data

Previously, we noted that interval and ratio scales of measurement produce real numbers that obey the standard rules of arithmetic (they can be added, subtracted, divided and multiplied). Interval and ratio measurements typically produce continuous data (e.g. weight, length, time, IQ). Each measurement has a specified level of precision reflecting the precision with which the measure has been taken; for example, the weight of a neonate could be measured as 4, 4.2, 4.18 or 4.183 kg depending on the accuracy of the scales that have been used to take the measurement and the precision needed in the clinical recording.

Grouped Frequency Distributions

When the continuous data are made up of a large number of varied measurements, it is useful to present the data as grouped frequency distributions. When drawing up a grouped frequency distribution, the following conventions should be taken into account:

Table 14.4

Score	F	Score	F	Score	F	Score	F
100	1	91	0	82	1	73	4
99	1	90	1	81	2	72	2
98	0	89	0	80	2	71	2
97	0	88	1	79	2	70	0
96	1	87	2	78	2	69	1
95	1	86	1	77	4	68	1
94	1	85	1	76	3	67	1
93	1	84	3	75	4	66	1
92	1	83	2	74	0		

1. The table of grouped frequency distributions should normally have no more than nine groups of values, otherwise it is too difficult to inspect. However, if too few groups are used, the meaning of the data is obscured, as varied measurements are combined into too few equivalent categories.

2. There should be equally sized class intervals between the groups (e.g. 1–5, 6–10, 11–15 rather than 1–3, 4–10, 10–20), the *width* of which is represented by i . In the first example, $i = 5$).
3. Individual scores within a given group ‘lose’ their precise identity. The midpoint of each class interval is taken to represent the value of the cases included in the group.

Example

On admission to hospital, patients are routinely weighed. You are asked to summarize the weights of 50 male patients who were admitted in your ward over a period of time. The weights (raw data) are as follows (to nearest kg):

75, 67, 76, 71, 73, 86, 72, 77, 80, 75, 80, 96, 93, 75, 73, 83, 81, 82,
73, 92, 81, 87, 76, 84, 78, 79, 99, 100, 88, 77, 71, 76, 75, 83, 66,
79, 95, 85, 77, 87, 90, 73, 72, 68, 84, 69, 78, 77, 84, 94

The steps in constructing a grouped frequency distribution are:

1. Organize the data into an ordered sequence, and find the frequency of each score (see [Table 14.4](#)).
2. Find the range of scores. The range is the difference between the highest and lowest score plus 1 (count on your fingers if you don’t believe this!). We add 1 to include the real limits for continuous data. In this case the range is $100 - 66 + 1 = 35$.
3. Decide on the width (i) of the grouped categories (technically we call these *class intervals*); i can be approximated by dividing the range by the number of groups or class intervals. In this instance if we decide on seven classes, i will be $35 \div 7 = 5.0$. When i is a decimal, it should be rounded up to the nearest whole number: here, $i = 5$. As stated earlier, the number of class intervals is arbitrary and will be chosen by the researcher, depending on the properties of the data. By convention, more than nine class intervals are rarely employed.

Table 14.5**Grouped Frequency Distribution of Patients' Weight in a Given Ward**

Class Interval	<i>F</i>
66–70	4
71–75	12
76–80	13
81–85	9
86–90	5
91–95	4
96–100	3
	<i>n</i> = 50

4. The next step is to determine the lowest class interval, and then list the limits of each class interval. Clearly, the lowest class interval must include the lowest score in the distribution.
5. Then, the frequency of scores is determined from each class interval and tabulated, as presented in [Table 14.5](#).

It is easier to understand the data by inspecting [Table 14.5](#) than by looking at the raw data. However, some precision in the data has been lost as somewhat different scores have been assigned into the same class intervals. The presentation of grouped data results in the loss of precision for the individual scores grouped together.

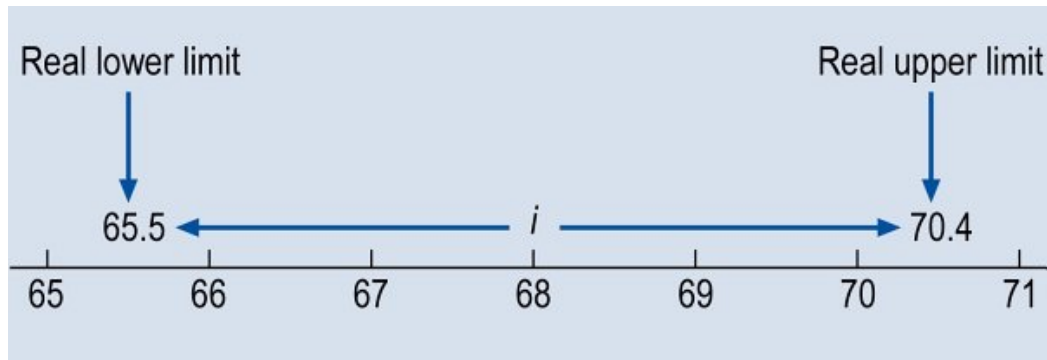


FIG. 14.4 Real limits of a class interval.

Graphing Frequency Distributions

The two common types of graphs used to graph frequency distribution of quantitative data are histograms and frequency polygons.

Histograms

A histogram resembles a bar graph but the bars are drawn to touch each other. The fact that the bars touch each other reflects the underlying continuity of the data that make up the bars. The height of the bars along the y -axis represents the frequency of each score or class interval plotted along the x -axis. With grouped data, the midpoint of each class interval becomes the midpoint of each bar, and the width of the bar corresponds to the real limits of each class interval.

For example, consider the lowest class interval, 66–70, in [Table 14.5](#). Because the data are continuous, the real upper and lower limits of the class interval are 65.5 and 70.4 ([Fig. 14.4](#)). Although all of the weights are given as whole numbers of kilograms, these will in most cases be the result of rounding off by the nursing staff to the nearest whole number. Thus someone who actually weighed 70.2 kg would have been recorded as weighing 70 kg and would fall into the 66–70 class. As [Fig. 14.4](#) shows, $i = 5$, and the midpoint of the class interval is 68.

Frequency Polygons

Any data that can be represented by a histogram can also be graphed as a frequency polygon. For this type of graph, a point is plotted over the

midpoint of each class interval, at a height representing the frequency of the scores. [Fig. 14.5](#) represents a histogram and a frequency polygon for the data in [Table 14.5](#).

Frequency polygons allow the reader to interpolate, that is, to estimate the frequency of values in between those actually measured or graphed. Of course, interpolation cannot be done for discrete data (see [Fig. 14.1](#)), as values between categories have no meaning. When a frequency polygon is plotted, it can take on a variety of shapes. The shapes which are of particular importance for frequency polygons are shown in [Fig. 14.6](#).

[Fig. 14.6A](#) represents a bell-shaped, normal distribution. It is symmetric in the sense that one-half is the mirror image of the other. The curve indicates that most of the scores fall in the middle, with relatively few scores toward either ‘tail’. [Fig. 14.6B](#) represents a negatively skewed distribution, with most of the scores being high and spreading out toward the lower end of the distribution. [Fig. 14.6C](#) represents a positively skewed distribution, with most of the scores being low, but with some scores spreading out towards the upper end of the distribution.

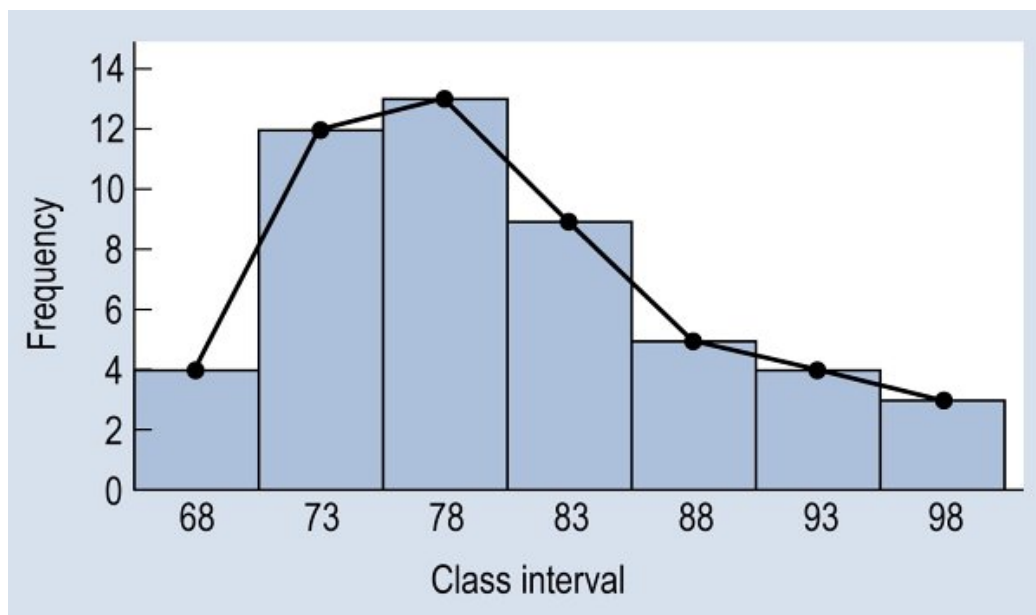
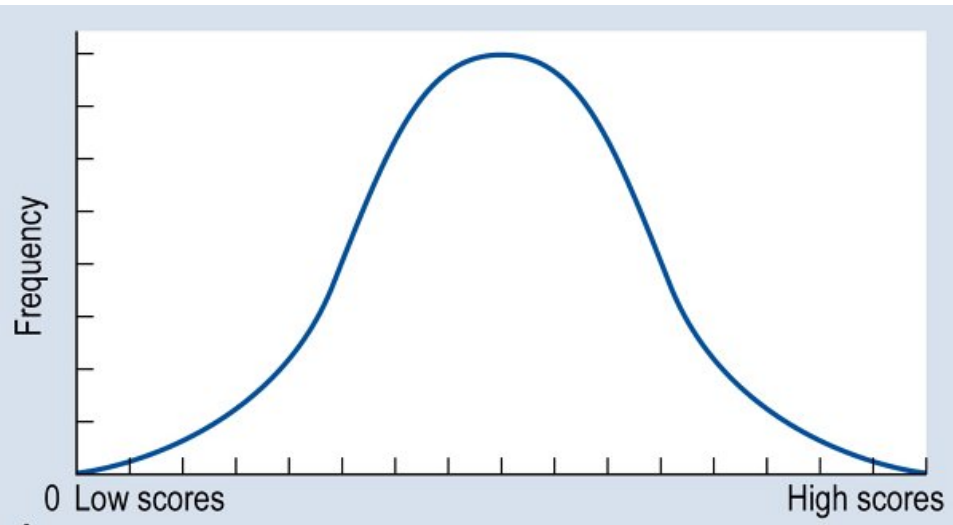
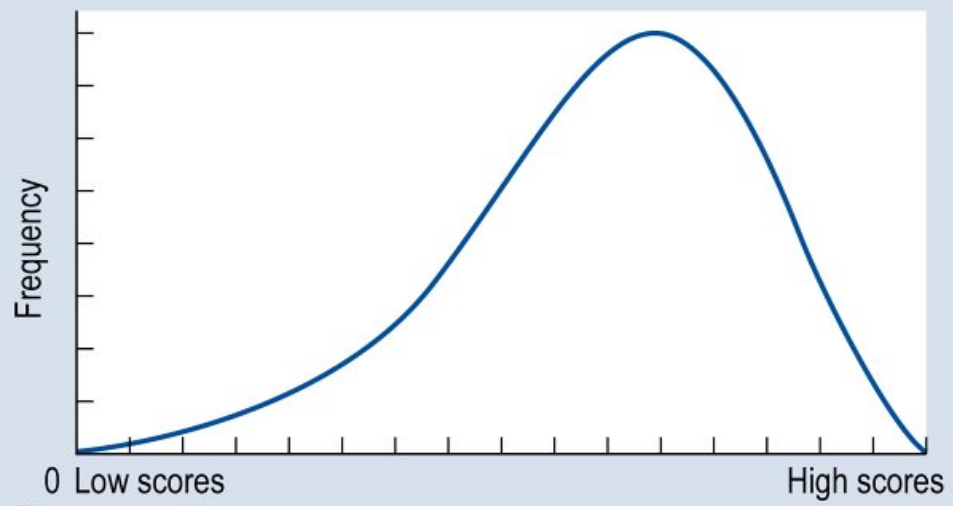


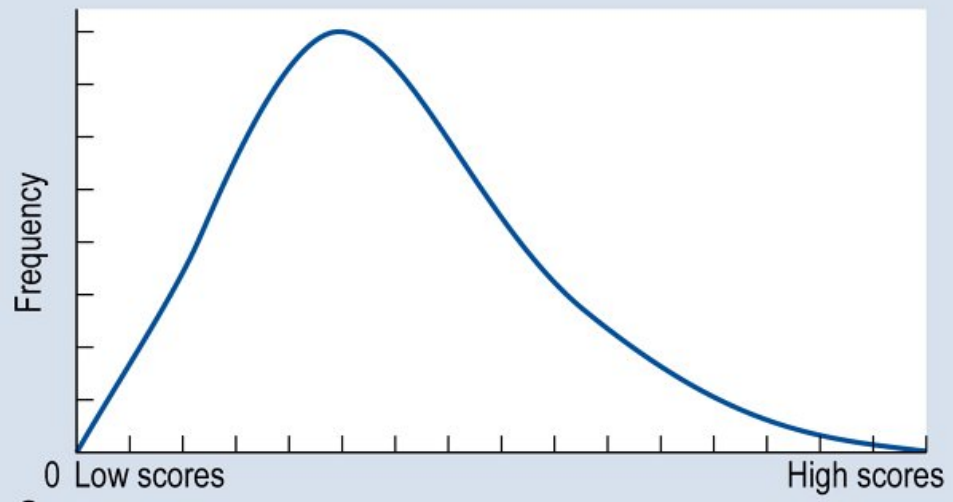
FIG. 14.5 Combined histogram and frequency polygon for the same data (see [Table 14.5](#)).



A



B



C

FIG. 14.6 Shapes of frequency polygons: (A) normal distribution; (B) negative skewing; (C) positive skewing.

An easy way to remember the direction of the skew is to consider the region where the ‘tail’ or portion of the graph with lower frequencies falls. For example, if the tail is located at the low score end of the x -axis, the graph is negatively skewed. The importance of the skew or a symmetric, normal distribution for data analysis will be discussed in subsequent sections.

Descriptive Statistics for Discontinuous Data

Once the data have been summarized in frequency distribution, it is often useful to make comparisons concerning the relative frequencies of scores falling into specific categories. As stated earlier, a statistic is a number generated by computations based on the raw data. The calculation of statistics is essential for ‘crunching’ the raw data into single numbers that represent the characteristics of the full data set. The following statistics are useful for understanding comparative trends in the data, and can be used for measurements on any scale: nominal, ordinal, interval or ratio.

Ratios

Ratios are statistics which express the relative frequency of one set of frequencies, A , in relation to another, B . The formula for ratios is:

$$\text{Ratio} = \frac{A}{B}$$

Therefore the ratio of males to females for the data presented in [Table 14.1](#) is:

$$\text{Ratio (males to females)} = \frac{10}{20} = 0.5$$

or

$$\text{Ratio (females to males)} = \frac{20}{10} = 2.0$$

Ratios are useful in the health sciences when we are interested in the distribution of illnesses or symptoms or the categories of participants requiring or benefiting from treatment. The ratio just calculated tells us about the relative frequency of gall bladder surgery for males and females.

Proportions

Proportions are statistics which are calculated by putting the frequency of one category over that of the total number in the sample or the population:

$$\text{Proportion of } A = \frac{A}{A + B}$$

Therefore the proportion of males in the sample represented in [Table 14.1](#) is:

$$\text{Proportion of males} = \frac{f(\text{males})}{n} = \frac{10}{30} = 0.33$$

Percentages

Proportions can be transformed into percentages, by multiplying by 100. Of course, this is how we obtained the values of the y-axis for [Fig. 14.2](#). To illustrate, patients scoring 5 (excruciating pain) in the control group:

$$\text{scoring 5} = \frac{5}{20} \times 100 = 25$$

The values of the pie chart (see [Fig. 14.3](#)) also involved such calculations.

Rates

When summarizing the results of epidemiologic investigations it often is useful to use this statistic to represent the level at which a disorder is present in a given population. The two rates which are commonly used in the health sciences are:

1. *Incidence rate*, which represents the number of new cases of a disorder reported within a time period:

$$\text{incidence rate} = \frac{\text{number of new cases of a disorder}}{\text{total population at risk of the disorder}} \times \text{base}$$

2. *Prevalence rate*, which represents the total number of cases suffering from a disorder:

$$\text{Prevalence rate} = \frac{\text{number of existing cases of a disorder}}{\text{total population at risk of the disorder}} \times \text{base}$$

Let us illustrate this equation by applying it to hypothetical data. Let us consider the condition herpes, a nasty little condition associated with the virus herpes simplex that attacks various parts (lips, etc.) of the body. Assume that an epidemiologist is interested in the spread of the condition in a given community.

1. Assume that all the population over the age of 15 years ($N = 1\,000\,000$) is at risk for herpes.
2. In 2005, there were 5000 new cases reported.
3. In 2005, there was a total (old and new active cases) of 15 000 known cases.

Here, substituting into the equation:

$$\text{Incidence rate for herpes} = \frac{5000}{1\,000\,000} \times \text{base}$$

0.005 × base

The statistic 0.005 is not seen as the best way to represent a rate.

Often, epidemiologists select a base to make the statistic more understandable. The base represents a number for transforming the rate. The base selected depends on the magnitude of the rate; conventionally a multiple of 10, such as 1000, 10 000 or 100 000 is selected. In this instance we select 1000 as the base. Therefore substituting into the equation, we obtain:

$$\text{Incidence rate for herpes simplex} \\ \frac{0.005 \times 1000}{1} = 5 \text{ per 1000}$$

$$\text{Prevalence rate for herpes simplex} \\ \frac{15000}{1000000} \times 1000 = 15 \text{ per 1000}$$

These statistics can be graphed. For example, we may wish to graph the incidence of herpes simplex in the community over a period of 5 years. The (fictitious) data are shown in [Table 14.6](#).

The graph of the time-series for the incidence over time ([Fig. 14.7](#)) provides a visual impression of the changing incidence rate of the problem in question.

Odds

Odds refer to a ratio of the number of cases favouring an event to the number of non-events. For example, the odds of throwing a six with a dice are:

$$\text{Odds} = \frac{\text{number of events favouring '1'}}{\text{number of non-events '2, 3, 4, 5, 6'}} \\ \frac{1}{5} \\ 0.2$$

Table 14.6

Incidence of Herpes

Year	Incidence of Herpes (per 1000)
1982	8
1983	15
1984	17
1985	16
1986	10

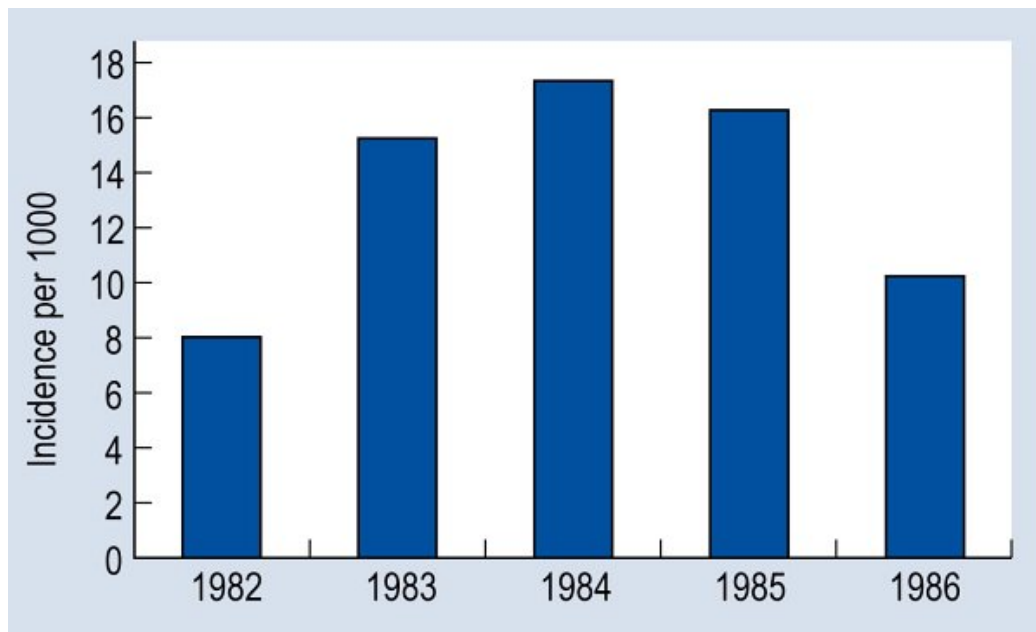


FIG. 14.7 Incidence rate of genital herpes 1982 to 1986.

In other words, the odds of throwing a six are one in five or 0.2. As you probably know, gamblers are very expert in using odds. These statistics are also important in health research as illustrated by the following example.

One way of assessing the effect of an intervention is to count the number of people who have, or who develop, a disease or health condition that the

intervention aimed at remediating. As a comparison, we also count the number of people who do not have or do not develop this disease or health condition. When there are two groups, this procedure produces dichotomous outcomes. The data for dichotomous outcomes can be summarized (see [Table 14.6](#)). Say that 200 obese, but otherwise healthy 55- to 65-year-old males are randomly assigned to two equally sized groups: the participants in the intervention group receive a carefully targeted exercise program, whereas participants in the control group maintain their usual lifestyle. The outcome measure is diagnosis of diabetes (yes or no) at the end of the trial. The hypothesis being tested is the odds of being diagnosed with diabetes in the intervention group will be reduced in comparison to the control group. The hypothetical findings are shown.

A number of values can be calculated from the data, shown in [Table 14.6](#), and these can be used to test the hypothesis regarding the benefits of the exercise program.

Table 14.7

Group	No Diabetes	Diabetes	Total
Intervention (exercise)	80 <i>a</i>	20 <i>b</i>	100
Control	60 <i>c</i>	40 <i>d</i>	100
Total	140	60	200

The proportion of people in each group who do not develop diabetes, is calculated as follows:

Intervention group:

$$\text{Proportion} = \frac{a}{a + b} = \frac{80}{100} = 0.8$$

Control group:

$$\text{Proportion} = \frac{c}{c+d} = \frac{60}{100} = 0.6$$

The proportion of people without diabetes is higher in the intervention group than in the control group.

Intervention group:

$$0.8 \times 100 = 80$$

Control group:

$$0.6 \times 100 = 60$$

The hypothetical results indicate that 80% of participants in the intervention group did not develop diabetes compared with 60% of participants in the control group. This outcome would support the efficacy of the exercise program.

The odds for each group of experiencing the adverse event is calculated as follows.

Intervention group:

$$\text{Odds} = \frac{a}{b} = \frac{80}{20} = 4.0$$

Control group:

$$\text{Odds} = \frac{c}{d} = \frac{60}{40} = 1.5$$

The odds of not developing diabetes in the intervention group are better than the odds for the control group. This is not the complete story; in [Chapter 20](#) we will look at how these results are used for determining the efficacy of the intervention. To do this, we will use a statistic referred to as the odds ratio (OR). To illustrate the concept of an OR for the present example:

$$\text{OR} = \frac{\text{odds intervention group}}{\text{odds control group}} = \frac{4}{1.5} = 2.7$$

In other words, participants had 2.7 times the odds of not developing diabetes if they had followed the exercise program.

Summary

We have outlined several techniques for organizing, tabulating and graphically presenting both discontinuous (nominal, ordinal) and continuous (interval, ratio) data. It was shown that raw data can be organized and tabulated as a frequency distribution, by counting the number of cases falling into specific categories or class intervals. Data composed of a large number of highly varied measurements were shown to be best presented by grouping the scores into class intervals.

Several techniques of graphing data have been discussed: bar graphs and pie charts for discontinuous data, and histograms and frequency polygons for continuous data. The possible shapes of frequency polygons were also examined. It was shown that discrete data grouped in frequency distributions can be represented as ratios, proportions, percentages, rates and odds.

Measures of Central Tendency and Variability

Introduction,
Measures of Central Tendency,
 The Mean or Average,
 The Mode,
 The Median,
Comparison of the Mode, Median and Mean,
Measures of Variability,
 The Range,
 The Average Deviation,
 The Variance,
 The Standard Deviation,
 The Semi-Interquartile Range,
Summary,

Introduction

In the previous chapter we examined how raw data can be organized and represented by the use of statistics to easily communicate and be

understood. The two types of statistics that are necessary for representing sample data are measures of *central tendency* and *variability/dispersion*.

Measures of central tendency are statistics or numbers expressing the most typical or representative scores in a data distribution. You are very likely already familiar with the mean or average, and others including the mode and median will be explained in this chapter. *Measures of variability* are statistics representing the extent to which scores are dispersed (or spread out) numerically. You are also likely to be familiar with ‘range’, which along with other measures of dispersion, including variance and standard deviation, will be explained in this chapter. The overall aim of this chapter is to explain the use of several types of measures of central tendency and variability commonly used in the health sciences and clinical practice. As most basic quantitative evidence arising from investigations is presented in terms of these statistics, it is essential to understand these concepts.

The aims of this chapter are to:

1. Discuss the selection and use of measures of central tendency.
2. Discuss the selection and use of measures of variability.
3. Outline the relationship between the skew of frequency distributions and the selection of appropriate descriptive statistics.

Measures of Central Tendency

The Mean or Average

You may already know the average (mean) of a group of values is obtained by adding all the scores and then dividing by the number of scores (e.g. the average of the heights for three family members who are 170 cm, 175 cm and 178 cm tall is found by adding 170 cm + 175 cm + 178 cm and dividing by 3. This yields an average result of 174.3 cm). We use special symbols to represent the average or mean. The mean ($\bar{x}$) or μ is defined as the sum of all the scores divided by the number of scores. The mean is, in fact, the arithmetic average for a distribution. The mean is calculated by the following equations:

$$\bar{X} = \frac{\sum x}{n} \text{ for a sample}$$

$$\mu = \frac{\sum \mu}{n} \text{ for a population}$$

where $\sum x$ = the sum of the scores, $\bar{X}$ = the mean of a sample, μ = the mean of a population, x = the values of the variable, that is, the different elements in a sample or population, and n or N = the number of scores in a sample or population. In statistical notation the sum or added values of a set of scores is represented by Σ (Greek sigma).

Table 15.1

Score	Real class interval	<i>F</i>	Cum <i>F</i>
1	0.5-1.4	2	2
2	1.5-2.4	5	7
3	2.5-3.4	4	11 *
4	3.5-4.4	3	14
5	4.5-5.4	3	17
		<i>n</i> = 17	

* Class interval containing median.

The formula simply summarizes the following calculation method:

To calculate the average or mean of a set of scores ($\bar{X}$), add together all the scores ($\sum x$) and divide by the number of cases (n).

Therefore given these sample scores:

2, 3, 5, 6, 7

To calculate the mean:

When n or N is very large, the average or mean is calculated with the previous formula but usually with the assistance of computers.

The Mode

When the data are categorical (the technical term for categorical data is *nominal*), the appropriate measure of central tendency is usually the mode. The mode is the most frequently occurring score or category in a study sample distribution. Therefore for the data shown in [Table 15.1](#), the mode is the ‘females’ category. The mode can easily be obtained by inspection of grouped data (with the largest group being the mode). As we shall see later, the mode also can be calculated for continuous data as well as discrete data.

The Median

With ordinal, interval or ratio-scaled data, central tendency also can be represented by the median. The median (Mdn) is the score that divides the distribution of the scores being studied into half: half of the scores in the study sample fall under the median, and half above the median. That is, if scores are arranged in an ordered array from highest to lowest or vice versa, the median would be the middle score. With a large number of cases and scores, it may not be feasible to locate the middle score simply by inspection. To calculate which is the middle score, we can use the formula $(n + 1)/2$, where n is the total number of cases in a sample. This formula gives us the number of the middle score. We can then count that number from either end of an ordered array.

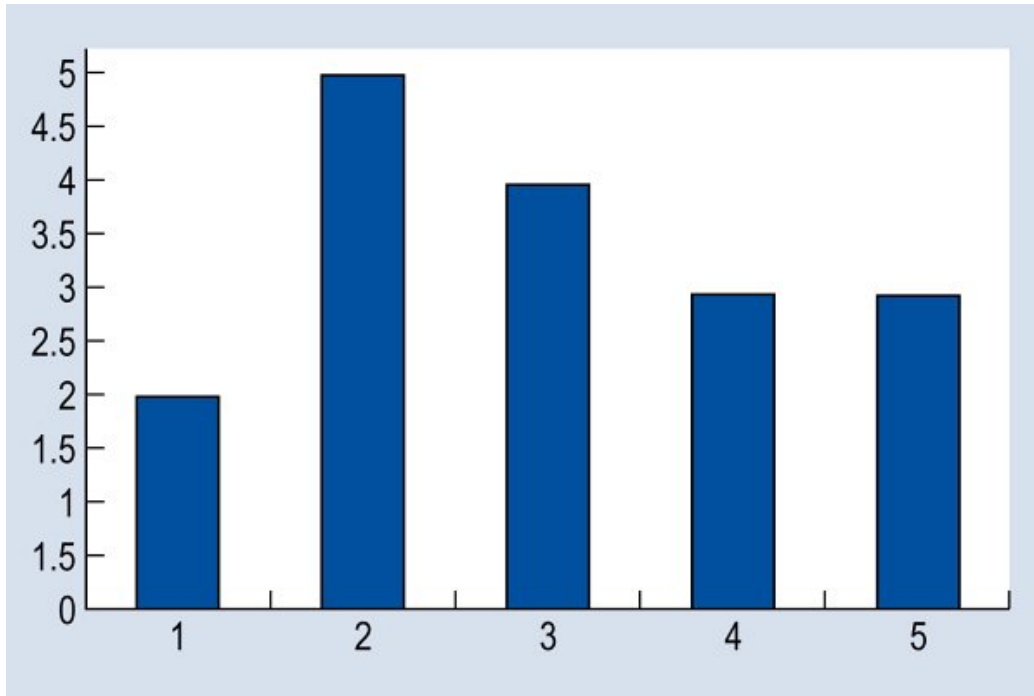


FIG. 15.1 Distribution of pain scores.

$$Mdn = X_L + i \left[\frac{\left(\frac{n}{2} \right) - cumf_L}{f_i} \right]$$

In general, if n (i.e. the number of cases) is odd, the median is the middle score; if n is even, then the median falls between the two centre scores. The formula $(n + 1) / 2$ is again used to tell us which score in an ordered array will be the median. For example:

5, 8, 9, 10, 28. Mdn = 9 (n is odd)

6, 17, 19, 20, 21, 27. Mdn = 19.5 (n is even)

For a grouped frequency distribution, the calculation of the median is a little more complicated. If we assume that the variable is continuous (e.g. time, height, weight or level of pain), we can use a formula for calculating the median. This formula (explained in detail later) can be applied to

ordinal data, provided that the variable being measured has an underlying continuity. For example, in a study of the measurement of pain reports we obtain the following data, where $n = 17$:

1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5

These data can be represented by a bar graph (Fig. 15.1).

Here we can obtain the mode simply by inspection. The mode = 2 (the most frequent score). For the median, we need the ninth score, as this will divide the distribution into two equal halves (see Table 15.1). By inspection, we can see that the median will fall into category 3. Assuming underlying continuity of the variable and applying the previously discussed formula, we have: where X_L = real lower limit of the class interval containing the median, i = width of the class interval, n = number of cases, $\text{cum } f_L$ = cumulative frequency at the real lower limit of the interval and f_i = frequency of cases in the interval containing the median.

Table 15.2

A		B	
X	F	X	F
1	16	1	8
2	1	2	7
3	1	3	6
4	7	4	4
	$n = 25$		$n = 25$

Substituting into the equation just presented:

$$Mdn = 2.5 + 1 \left[\frac{\left(\frac{17}{2} \right) - 7}{4} \right] = 2.5 + 1 \left[\frac{1.5}{4} \right] = 2.875$$

Comparison of the Mode, Median and Mean

The mode can be used as a measure of central tendency for any level of scaling. However, because it only takes into account the most frequent scores, it is not generally a satisfactory way of presenting central tendency because it does not take into account scores from the whole of the distribution. For example, consider two sets of scores, A and B, shown in [Table 15.2](#). It can be seen, either by inspection or by sketching a graph, that the two distributions A and B are quite different, yet the modes are the same (i.e. 1).

The median divides distributions into two equal halves, and is appropriate for ordinal, interval or ratio data. For interval or ratio data, however, the mean is the most appropriate measure of central tendency. The reason for this is that, in calculating this statistic, we take into account all the values in the study sample. In this way, it gives the best representation of the average score. Clearly, it is inappropriate to use the mean with nominal data, as the concept of *average* does not apply to discrete categories. For example, what would be the average of 10 males and 20 females?

There is some justification for using the median as a measure of central tendency when the variable being measured is continuous. However, the mean is almost always used and preferred. Alternatively, when a distribution is highly skewed (asymmetrical), the median might be more appropriate than the mean for representing the ‘typical’ score. Consider the distribution of ratings roses:

2, 2, 2, 5, 7, 8, 9

mode = 2

median = 5

$\bar{x} = 5$

Let us change the 9 to 44:

2, 2, 2, 5, 7, 8, 44

mode = 2

median = 5

$\bar{x} = 10$

Clearly, the median and the mode are less sensitive to extreme scores, whereas the mean is pulled towards extreme scores. This might be a disadvantage. For example, there are seven people working in a small factory, with the incomes per week presented here:

\$100, \$200, \$200, \$300, \$400, \$400, \$1900

Median = \$300

$\bar{x} = 500$

The distribution of wages is highly skewed by the high income of the owner of the factory (\$1900). The mean, \$500, is higher than six of the seven scores; it is in no way typical of the distribution. In cases like this the median is more representative of the distribution.

Fig. 15.2 illustrates the relationships between the skew of frequency distributions and the three measures of central tendency discussed in this chapter. We should remember that the mode will always be at the highest point, the median will divide the area under the curve into halves, and the mean is the average of all the scores in the distribution. Also, the greater the skew in distribution, the more the measures of central tendency are likely to differ.

Measures of Variability

We have seen that a single statistic can be used to describe the central tendency of a frequency distribution. This information is insufficient to characterize a distribution; we also need a measure of how much the scores are dispersed or spread out. The variability of discrete data is of little relevance, as the degree of variability will be limited by the number of categories defined by the investigator at the beginning of measurement.

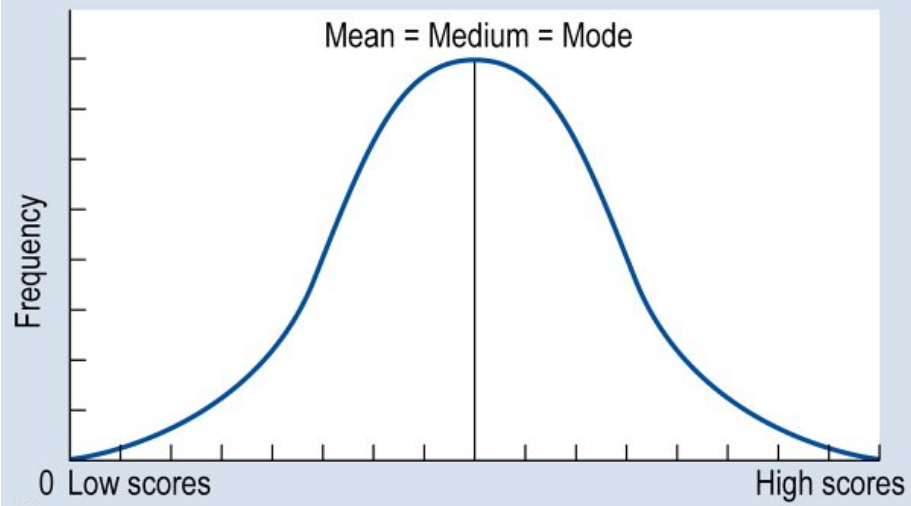
Consider the next two hypothetical distributions representing the IQs of two groups of intellectually disabled children:

Group A: 45, 50, 55, 60, 60, 70, 80

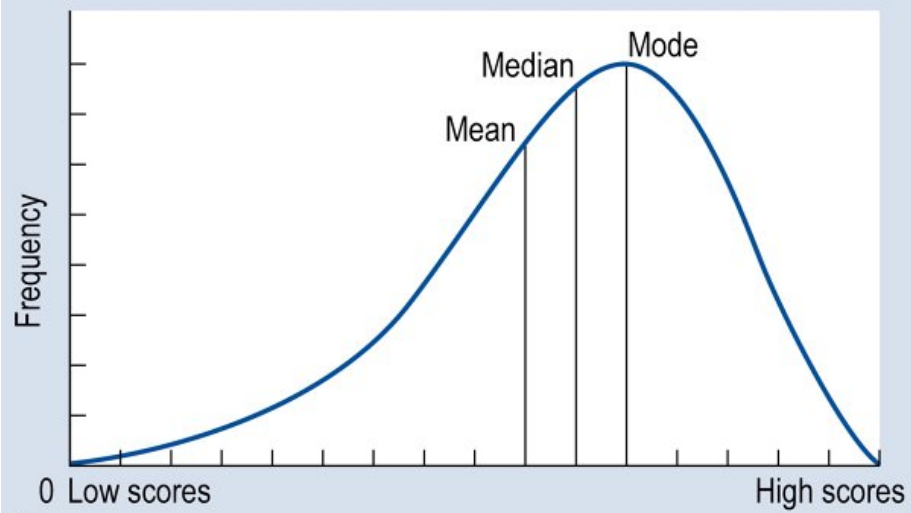
Group B: 57, 58, 59, 60, 61, 62, 63

It is evident that, although $\bar{X}_A = \bar{X}_B = 60$, the variability of the scores of group A is greater than that of group B. Insofar as IQ is related to the activities appropriate for these children, group A will provide a greater challenge to the therapist working with the children.

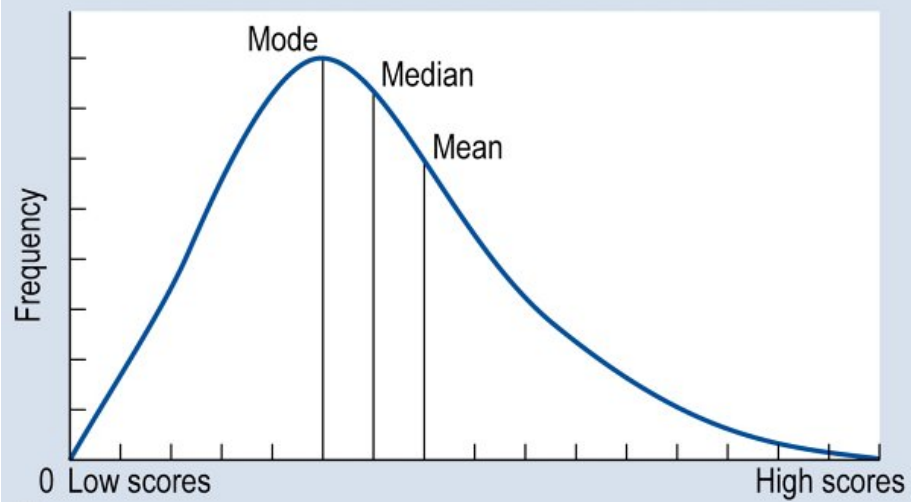
The three statistics commonly used to indicate the numerical value of variability are the *range*, the *variance* and the *standard deviation*.



A



B



C

FIG. 15.2 Measures of central tendency in (A) normal distribution, (B) negative skewing, (C) positive skewing.

The Range

The range is the difference between the highest and lowest scores in a distribution. As we mentioned, given the IQ data just presented, the ranges are:

$$\text{Group A: } 80 - 45 + 1 = 36$$

$$\text{Group B: } 63 - 57 + 1 = 7$$

Although the range is easy to calculate, it is dependent only on the two extreme scores. In this way, we are not representing the typical or average variability of the scores in the study sample. That is, the range might be distorted by a small number of extreme atypical scores or *outliers*. Consider, for instance, the differences in the range for the data given. The earlier example of the distribution of wages shows that just one outlying score in a distribution has an enormous effect on the range (\$1900 – \$100 = \$1800). Obviously, some measure of average variability would be a preferable way to represent the variability.

Table 15.3

<i>X</i>	
57	– 3
58	– 2
59	– 1
60	0
61	+ 1
62	+ 2
63	+ 3

The Average Deviation

At first consideration, a convenient measure of variability might be the average deviation about the mean. Consider group B shown previously.

Here $\bar{x} = 60$. To calculate the average variability about the mean, we subtract the mean from each score, sum the individual deviations, and divide by n , the number of measurements ([Table 15.3](#)). Therefore:

$$\begin{aligned} \sum (x - \bar{x}) \\ = (-3) + (-2) + (-1) + (0) + (1) + (2) + (3) = 0 \end{aligned}$$

This is a general result; the sum of the average deviations about the mean is always zero. You can demonstrate this for the average deviation of group A. So actually, the average deviation is kind of useless because it has the same value of zero no matter how spread out the scores are. This is a problem because you will always get the same zero result using this method. The problem can be solved by squaring the deviations, as the square of negative numbers is always positive, and this gives a much better measure of variability. This statistic is called the *sums of squares* (SS) and is always a positive number. This squaring procedure of deviations from the average or mean leads to a new statistic called the variance.

The Variance

The variance (σ^2 or s^2) is defined as the sum of the squared deviations about the mean divided by the number of cases:

Table 15.4

x	$x - \bar{x}$	
57	-3	9
58	-2	4
59	-1	1
60	0	0
61	+1	1
62	+2	4
63	+3	9

For a population: $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$

For a sample: $s^2 = \frac{\sum (x - \bar{X})^2}{n}$

Divide by $n - 1$ when calculating the variance for a sample, when we use s^2 as an estimate of population variance. Dividing by n results in an estimate which is too small, given that a degree of freedom has been lost calculating $\bar{X}$.

For the earlier IQ example, the variance is calculated as shown in [Table 15.4](#).

Substituting into the formula:

$$\Sigma^2 = \frac{\sum (x - \bar{X})^2}{n - 1} = \frac{28}{6} = 4.67$$

The problem with this measure of variability is that the deviations were squared. In this sense, we are overstating the real spread of the scores. So, we can restore them by taking the square root of the variance and we then arrive at the most commonly used measure of variability for continuous data: the standard deviation. When presenting summary tables of continuous data, it is a universal procedure to present the mean and the standard deviation of each variable. So, you need to understand how to calculate the mean and the standard deviation.

The Standard Deviation

The standard deviation (σ or s) is defined as the square root of the variance:

$$\Sigma = \sqrt{\Sigma^2} = \frac{\sqrt{\sum (x - \mu)^2}}{N}$$

$$s = \sqrt{s^2} = \frac{\sqrt{\sum (x - \bar{X})^2}}{n - 1}$$

Therefore the standard deviation for group B is:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

The size of the standard deviation reflects the spread or dispersion of the frequency distribution of the data. Clearly, the larger the values of σ or s , the relatively more spread out the scores are in a distribution. Calculation of the variance or the standard deviation is extremely tedious for large n using the method shown here. Statistics texts provide a variety of calculation formulae to derive these statistics. However, the accessibility of computer-based statistical packages for analysing data in research and administration makes it superfluous to discuss these calculation formulae in detail, as almost all data are analysed using computer-based statistical packages.

The Semi-Interquartile Range

We have seen previously that, if we are summarizing ordinal data, or interval or ratio data that is highly skewed (i.e. the frequency distribution is lopsided and not symmetric), then the median is the appropriate measure of central tendency. The statistic called interquartile and semi-interquartile range is used as the measure of dispersion when the median is the appropriate measure of central tendency for a distribution. The interquartile range is the distance between the scores representing the 25th (Q_1) and 75th (Q_3) percentile ranks in a distribution.

It is appropriate to define what we mean by percentiles (sometimes called *centiles*). The percentile or centile rank of a given score specifies the percentage of scores in a distribution falling up to and including the score. As an illustration, consider [Fig. 15.3](#), in which:

- 25% of cases or scores fall up to and including Q_1 .
- 50% of cases or scores fall up to and including the median (Q_2).
- 75% of cases or scores fall up to and including Q_3 .
- 25% of cases or scores fall above Q_3 .

The distances A and B represent the distances between the median and Q_1 and Q_3 . When a distribution is symmetric or normal, the distances A and B will be equal. However, when a distribution is skewed (i.e. lopsided or asymmetric), the two distances will be quite different. The semi-interquartile range (sometimes called the *quartile deviation*) is half of the distance between the scores representing the 25th (Q_1) and 75th (Q_3) percentile ranks in a distribution. Let us look at an example. If we have a sample where $n = 16$ and the values of the variable are:

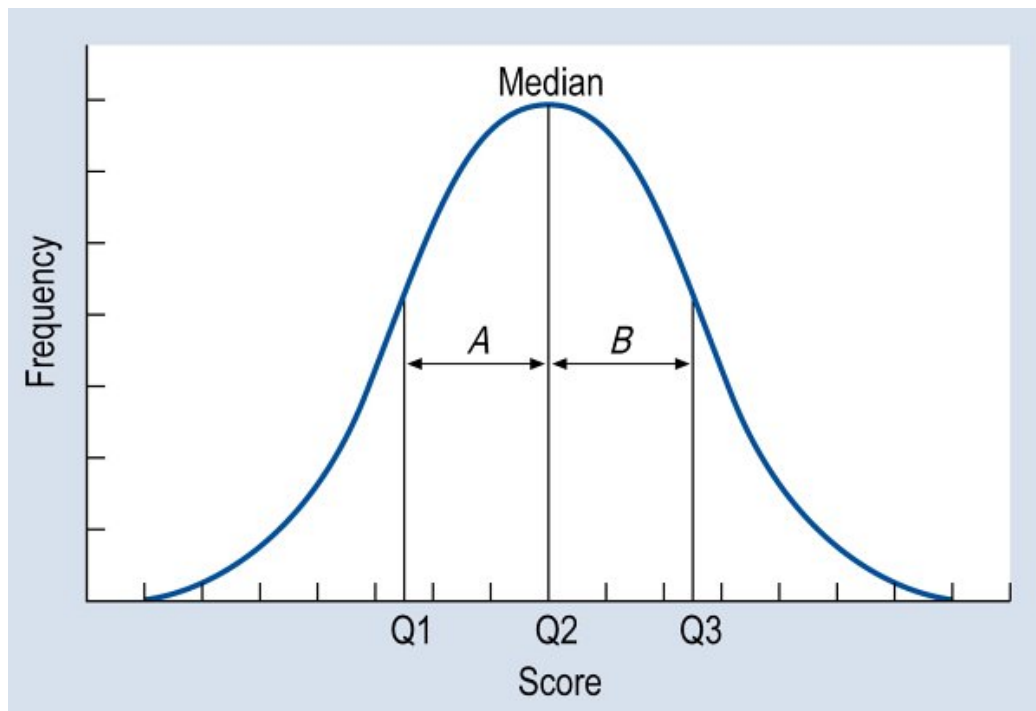


FIG. 15.3 Interquartile ranges.

1, 5, 7, 7, 8, 9, 9, 10, 11, 12, 13, 15, 19, 20, 20, 20

clearly, a frequency distribution of these data is not even close to normal as the distribution is not symmetric and the mode is at the maximum value. Therefore the median is selected as the appropriate measure for central tendency, and we should use the interquartile range as the measure of dispersion.

Looking at the data, we find:

1 5 7 7 * 8 9 9 10 † 11 12 13 15 ‡ 19 20 20 20

where * denotes the 25th centile, † denotes the 50th centile and ‡ denotes the 75th centile.

The score that cuts off the first 25% of scores (25th centile) is the first quartile (Q_1). Since we have $n = 16$, Q_1 will cut off the first four scores (25% of 16 is 4). Thus:

$$Q_1 = \frac{7 + 8}{2} = 7.5$$

The third quartile (Q_3) is the score which cuts off 75% of the scores. As $n = 16$, Q_3 will cut off 12 scores (75% of 16 is 12).

$$Q_3 = \frac{15 + 19}{2} = 17$$

Therefore the semi-interquartile range is:

$$\frac{Q_3 - Q_1}{2} = \frac{17 - 7.5}{2} = 4.75$$

The larger the semi-interquartile range, the more the scores are spread out about the median.

Summary

In this chapter, we discussed two essential statistics for representing frequency distributions: measures of central tendency and variability. The measures of central tendency outlined were the mode and median for discrete data, and the mean for continuous data. Measures of variability were shown to be the range, average deviation, variance and standard deviation. These statistics are appropriate for crunching the data to the point that the distribution of raw data can be meaningfully represented by only two statistics. That is, the raw data representing the outcome of investigations or clinical measurements are expressed in this manner. We have seen that the mean and the standard deviation are most appropriate for interval or ratio data. The median and semi-interquartile range are used when the data are measured on an ordinal scale, or when interval or ratio data are found to have a highly skewed (lopsided or asymmetrical) distribution. The mode represents the most frequent scores. The contents of the chapter focused on the use and meaning of these concepts, rather than stressing calculations involved. These calculations are obtained by using computer based statistical software packages. In [Chapter 16](#), we discuss the application of the mean and standard deviation for relating specific scores to an overall distribution.

Standard Scores and Normal Distributions

Introduction,
Standard Scores (z-Scores),
Normal Distributions,
 The Standard Normal Curve,
Calculations of Areas Under the Normal Curve,
 Critical Values,
Standard Normal Curves for the Comparison of
Distributions,
Summary,

Introduction

In this chapter we will discuss standard distributions. Standard scores represent the position of a score or measurement in relation to an overall set of scores. Standard distributions are also useful for comparing scores from different sets of measurements. Standard scores are used in both clinical practice and research in the health sciences. In clinical practice, a patient's score often is compared with a known distribution to determine whether the result is 'normal' (i.e. falls within a 'normal' range of values) or is abnormal

(is unusually low or high). Measurements such as blood pressure or cholesterol levels are compared with a distribution of scores from the general population or a subgroup (e.g. women or men) to interpret the patient's result. Normal blood pressure is considered by the American Heart Association (2017) to be <120 mm Hg systolic and <80 mm Hg diastolic. Normal blood sugar (HbA1c glycated haemoglobin) is considered to be <42 mmol/mol, or 6%.

The aims of this chapter are to:

1. Define 'standard' scores.
2. Describe the characteristics of normal and standard normal curves.
3. Show how standard normal curves can be used for calculating percentile ranks.
4. Show how standard normal curves can be used to compare scores from different distributions.

Standard Scores (z-Scores)

Consider this example: infant A walked unaided at the age of 40 weeks, whereas infant B is 65 weeks old but still cannot walk. What sense can we make of these measurements? Could infant B need further clinical investigation to determine if he or she has some neurologic abnormality? The fact that infant B is unable to walk at the age of 65 weeks is not very informative in the absence of additional information about how this compares with norms for other children. However, say that it is known that the distribution of walking ages is such that the population average or mean is $\mu = 50$ weeks and the population standard deviation is $\sigma = 5$ weeks. Assuming that the frequency distribution of first walking ages is normal, the frequency polygon representing the population would look something like that shown in [Fig. 16.1](#).

In this instance, infant B's score is clearly and substantially higher than the mean or average. In fact, we can see the infant's score at this point of time was 3 standard deviations over ($+3 \times 5$) the mean (i.e. $65 = 50 + (3 \times 5)$). In contrast, infant A began walking earlier than the mean (average), his or her score of 40 being 2 standard deviations below ($-2 \times 5 = 10$) the mean (which is 50). In general, any 'raw' score in a frequency distribution can be

described in terms of its distance from the mean or average. The process of transforming a score into a measurement based on its distance from the mean in standard deviations is called *standardizing* the score. Such ‘transformed’ scores are called z-scores or standard scores.

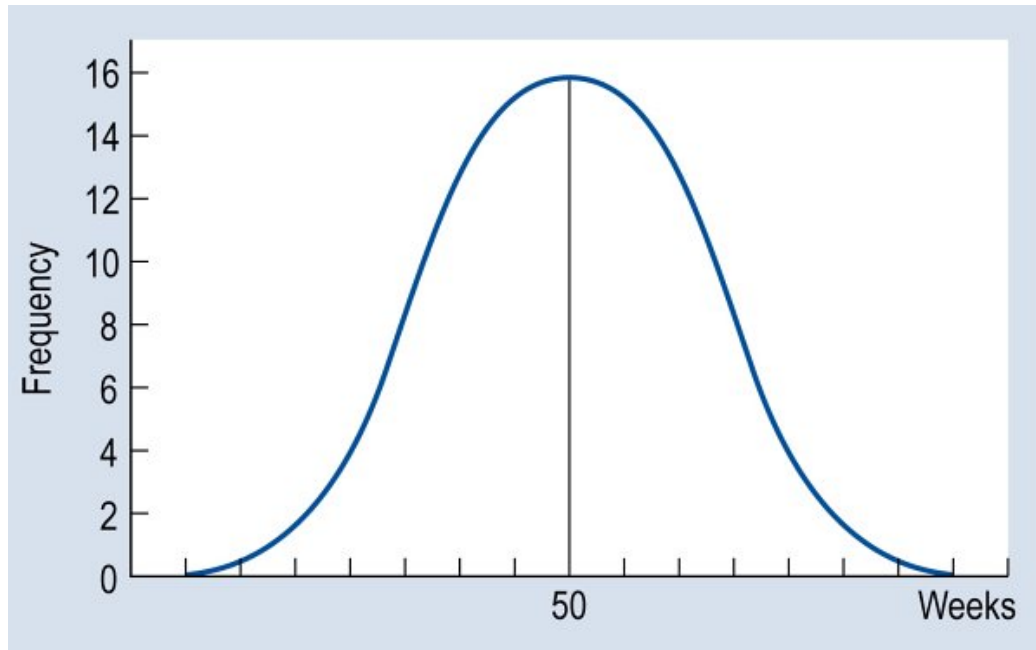


FIG. 16.1 Age at which children walk unaided.

A z-score represents how many standard deviations a given raw score is above or below the mean. The equation for transforming specific raw scores into z-scores is given as:

$$z = \frac{x - \mu}{\sigma} \text{ (for a population)}$$
$$z = \frac{x - \bar{X}}{s} \text{ (for a sample)}$$

For this equation, x is the raw score, $\bar{x}$ or μ is the mean of the distribution from which the score was drawn and s or σ is the standard deviation of the distribution. That is, when we know the mean and standard deviation of a distribution, we can transform any raw score into a z- (standard) score.

Conversely, when the z-score is known, we can use the previous equations to calculate the corresponding raw scores.

In the example just given, the z-scores corresponding to the infants' raw scores are:

$$\begin{aligned}\text{Infant A z score} &= \frac{X - \mu}{\sigma} = \frac{40 - 50}{5} = -2 \\ \text{Infant B z score} &= \frac{X - \mu}{\sigma} = \frac{65 - 50}{5} = +3\end{aligned}$$

These calculations support our previous observations that A's score was 2 standard deviations below the mean and B's score was 3 standard deviations above the mean. In other words, infant A walked very early and infant B was a very late starter. The particular value of standardizing scores for understanding clinical or research evidence will be discussed in the context of the concepts of normal and standard normal distributions. The standard score is a very useful value because it reveals the underlying meaning of the raw score.

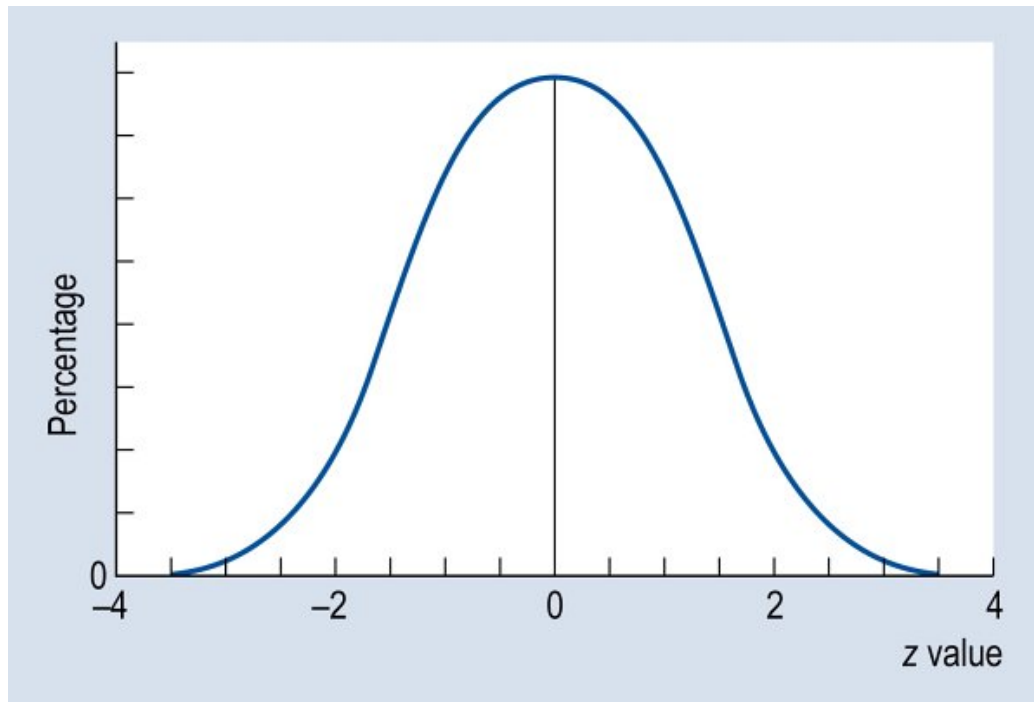


FIG. 16.2 Standard normal distribution.

Normal Distributions

Many variables measured in the biological, behavioural and clinical sciences are approximately ‘normally’ distributed (i.e. they have a characteristic ‘bell-shaped’ curve). What is meant by a normal distribution is illustrated by the normal curve (Fig. 16.2), which is a frequency polygon representing the theoretical distribution of population scores. We assume here that the variable x has been measured on an interval or ratio scale and that it is a continuous variable such as weight, height or blood pressure.

The normal curve has the following characteristics:

1. It is symmetric about the mean, so that equal numbers of cases fall above and below the mean (the value of the mean = the value of the median = the value of the mode).
2. Relatively few cases fall into the high or low values of x . Most of the cases fall close to the mean. (For the theoretical normal distribution, the arms of the curve do not intersect with the x -axis, allowing for a few extreme scores.)

3. The precise equation for the normal curve was discovered by the mathematician Gauss, so it is sometimes referred to as a Gaussian curve (i.e. Gaussian is a synonym for ‘normal’).

We need not worry about the actual formula for the normal curve. Rather, the point is that, given that the functional relationship between f and x is known, from mathematics we can use integral calculus to calculate areas under the curve for any value of x . All normal curves have the same general mathematical form; whether we are graphing IQ or weight, the same bell shape will appear. The only differences between the curves are the mean value and the amount of variation. This is why the mean and the standard deviation provide us with important information about any particular normal distribution. Note that it is unlikely that any real data are exactly normally distributed. Rather, the normal distribution is a mathematical model that is useful for representing real distributions.

The Standard Normal Curve

If we transform the raw scores of a variable into z -scores and then plot the frequency polygon for the distribution, we will have a standard curve. If the original distribution was normal, then the frequency polygon will be a standard normal curve. Standard normal curves are identical regardless of the nature of the original variables.

By transforming raw scores into z -scores, we are adjusting for differences in means and standard deviations, which are the only things that distinguish between nonstandardized normal curves. It is convenient to do this because it means we only need to have one set of statistical tables to perform various useful calculations.

The standard normal curve has the following additional properties:

1. The mean is always 0. For the previous example, the z -score corresponding to $\mu = 50$ (as in the ‘infants’ walking age’ example) is:

$$z = \frac{50 - 50}{5} = 0$$

2. The mean = median = mode, as the curve is symmetric.
3. Standard deviation above the mean is:

$$z = \frac{55 - 50}{5} = 1$$

4. It is assumed that the *total area under the curve adds up to 1.00*.
Since the normal curve is symmetric, 0.5 of the area falls above $z = 0$ and 0.5 falls below $z = 0$. This is another way of saying that 50% of the total cases fall below the mean, and 50% of the cases fall above the mean (which is equal to the median).
5. More generally, we can use appropriate *statistical tables* to estimate the area under the standard normal curve for any given z -scores.
These areas are available in table form (see Appendix A) so that for any value of z we can read off the corresponding area.
6. The area under the curve between any two points is directly proportional to the percentage of cases falling above, below or between those two points. We can use the standard normal curve to calculate the percentage of scores falling between any specified two scores.

In the next section, we will examine the use of the table of areas under the standard normal curve to understand the meaning of measurements in relation to distributions.

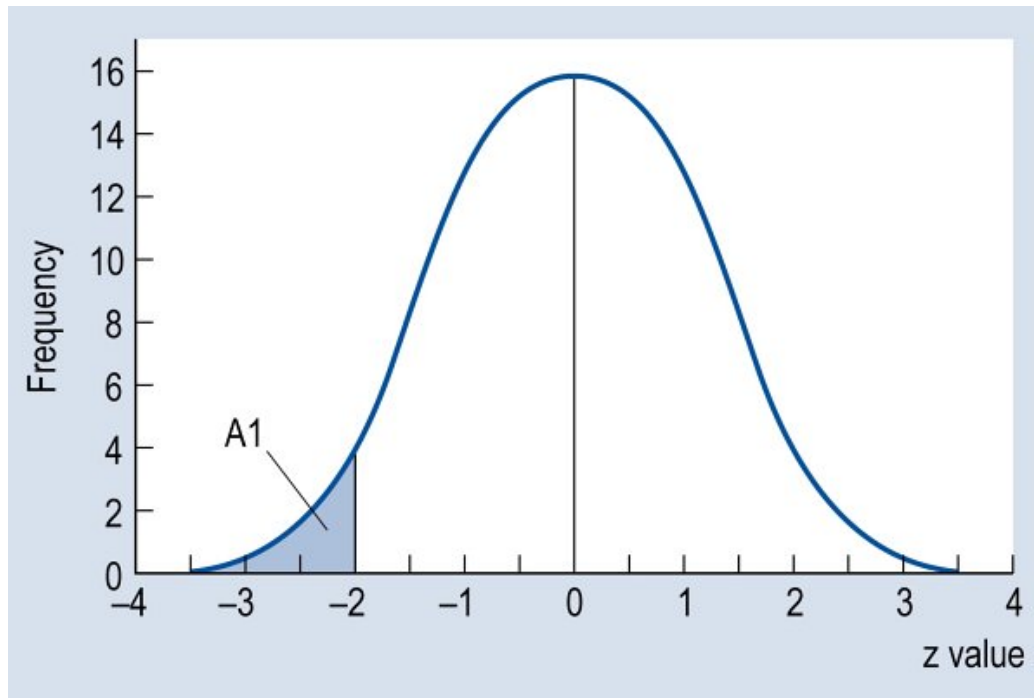


FIG. 16.3 Area (A1) corresponding to $z = -2$.

Calculations of Areas Under the Normal Curve

We have already examined the concept of percentile or centile ranks. The normal curve is useful for evaluating the percentile rank of scores in normal distributions. Appendix A gives the proportion of areas under the standard normal curve which lies:

- between the mean and a given z -score and
- beyond the z -score.

Since normal distributions are symmetric, the same proportions are also true for the area between the mean and any negative z -score. Only the positive values are given in Appendix A.

Let us see how we can use this information to estimate the percentile ranks of the two infants' walking ages. We have shown previously in our hypothetical example that, for infant A, $z = -2$. Let us now turn to Appendix

A. In going down the column of z -scores, we find that the area corresponding to $z = 2.00$ is 0.4772 (between) and 0.0228 (beyond).

We know that the area A1 under the curve in [Fig. 16.3](#) must be:

Shaded portion = $0.5000 - 0.4772$
0.0228 (as half the scores fall under the mean)

This proportion can be expressed as a percentage, so that 2.28% of the cases in the distribution fall below $z = -2$. We have defined percentile rank for a score as the percentage of cases in a distribution falling up to and including a specific score. Therefore the percentile rank for infant A's walking is 2.28%. Of all children, only 2.28% learn to walk as early as or earlier than infant A. Clearly, he is doing well.

What then is the percentile rank for infant B's performance? As you remember, $z = +3$. Looking up the area corresponding to $z = 3.00$ in Appendix A we find the area (A1 in [Fig. 16.4](#)) is equal to 0.4987. Therefore the proportion of scores falling up to and including $z = +3$ is $0.5 + 0.4987 = 0.9987$.

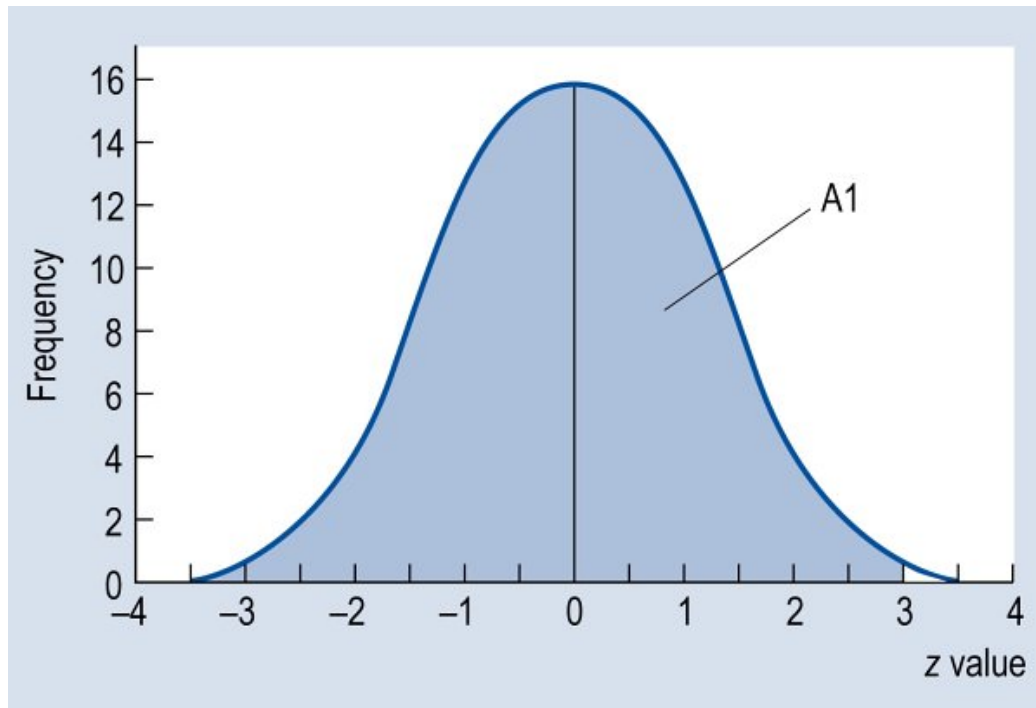


FIG. 16.4 Area (A1) corresponding to $z = +3$.

Expressing this finding as a percentage, we find that 99.87% of children learn to walk by the age of 65 weeks. As we said earlier, infant B is still not walking. Perhaps further clinical tests are indicated, although we should keep in mind that an unusual or extreme score is not necessarily indicative of pathologic states.

In paediatrics, growth charts showing the percentiles of children with certain physical characteristics are used widely to monitor the progress of children (see WHO, 2006). And many clinical standard values use similar charts showing the extent to which the patient's results deviate from the healthy normal values for clinical parameters.

Critical Values

We can also use these principles to work the other way by determining the raw scores corresponding to areas under the normal curve in Appendix A. For example, let's say it is recommended that the slowest 5% of infants (i.e. the 5% with the highest age first walked scores) are offered some special exercises in learning to walk. What would be the age at which the exercises should be offered, should the child not be walking? The key here is to

determine the score that corresponds to the slowest 5% (5th percentile) of the distribution. This can be represented as shown in Fig. 16.5.

From Fig. 16.5 we can see that we need to discover the z -score that corresponds to an area of 0.45 above the mean. By consulting the normal distribution table (see Appendix A), it can be seen that the corresponding z -score is $z = 1.65$. This is a critical value for the statistic in defining an area.

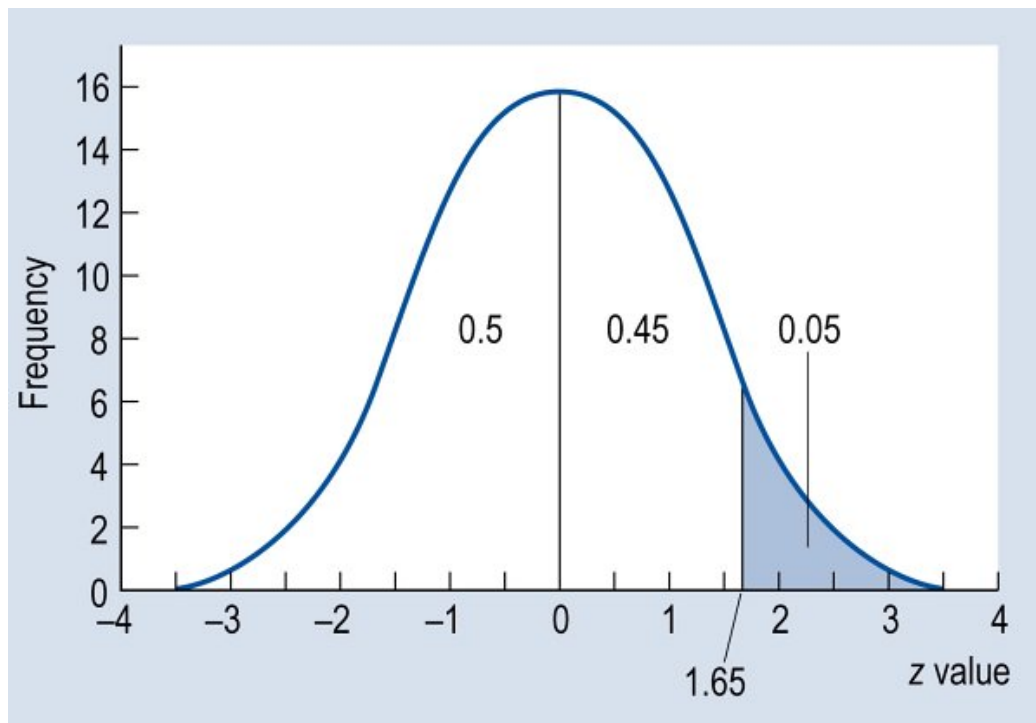


FIG. 16.5 Determining z -score of 95th percentile.

Given the z -score, we can calculate the corresponding raw score from the formula:

$$z \text{ score} = \frac{x - \mu}{\sigma}$$

We know that $z = 1.65$ so $1.65 = \frac{x - 50}{5}$

Multiply both sides by 5 and we get $16.5 \text{ times } 5 = x - 50$

So therefore $x = (5 \text{ times } 1.65) + 50 = 58.25$

That is, if the slowest 5% are thought to be in need of help, then children somewhat older than 58 weeks old and still not walking would be recommended for the remedial exercises. Of course, we can use the tables for reading off the z -scores corresponding to any specified area or percentage. It must be emphasized that if a distribution is *not* approximately normal (e.g. markedly skewed or bimodal), then the critical values just discussed do not exactly apply. It is always appropriate to check if a distribution is approximately normal or not before conducting statistical analysis. Most computer-based statistical software packages have this as an inbuilt feature. Actually, unless the distribution is seriously skewed or asymmetric the standard tables are a good approximation.

Standard Normal Curves for the Comparison of Distributions

One of the uses of standard distributions is that we can compare scores from entirely different distributions. For example, if a student scored 63 on test A, and 52 on test B, on which test did the student do better? If we define ‘better’ as solely in terms of raw scores, then clearly the student did better on test A. However, test A might have been easier than test B, so that if the overall performances of all students on the tests are taken into account, the student’s relative performance might be better on test B. Say:

Test A: $\bar{X} = 65$ and $s = 8$
Test B: $\bar{X} = 40$ and $s = 8$

Therefore using the formula for calculating z -scores and looking up the corresponding areas in Appendix A (do this yourself) we obtain the results shown in [Table 16.1](#). Thus the student performed better on test B, by scoring higher than 93% of other students sitting the test.

This example illustrates that in some circumstances the meaning of specific scores has to be interpreted against ‘standards’.

Another use of standard distributions is in interpreting the meaning of the results of investigations in the health sciences. Let us examine a hypothetical example. An investigator measured levels of blood cholesterol in a sample of 300 adults who are meat eaters, and 100 adults who are vegetarians. The results of the investigation are summarized in [Table 16.2](#).

Now, imagine that you are a clinician working with patients with cardiac disorders and you are interested in these two questions:

1. Approximately what percentage of vegetarians had blood cholesterol levels greater than the average meat eater?
2. Approximately what percentage of meat eaters had blood cholesterol levels lower than that of the average vegetarian?

The percentage of cases of vegetarians with blood cholesterol >0.6 (the mean for the meat eaters) is represented by area A1 in [Fig. 16.6](#).

$$z = \frac{0.6 - 0.4}{0.1} = 2$$

Area A1 = 0.0228

Therefore approximately 2.3% of vegetarians had blood cholesterol levels higher than the average for meat eaters.

Table 16.1

Scores for Example in Text

Raw scores	z-Scores	Percentile ranks
$x = 63$	-0.25	40.1
$x = 52$	+ 1.50	93.3

Table 16.2

Blood Cholesterol

	Mean blood cholesterol (mg/mL)	Standard deviation
Meat eaters	0.6	0.15
Vegetarians	0.4	0.1

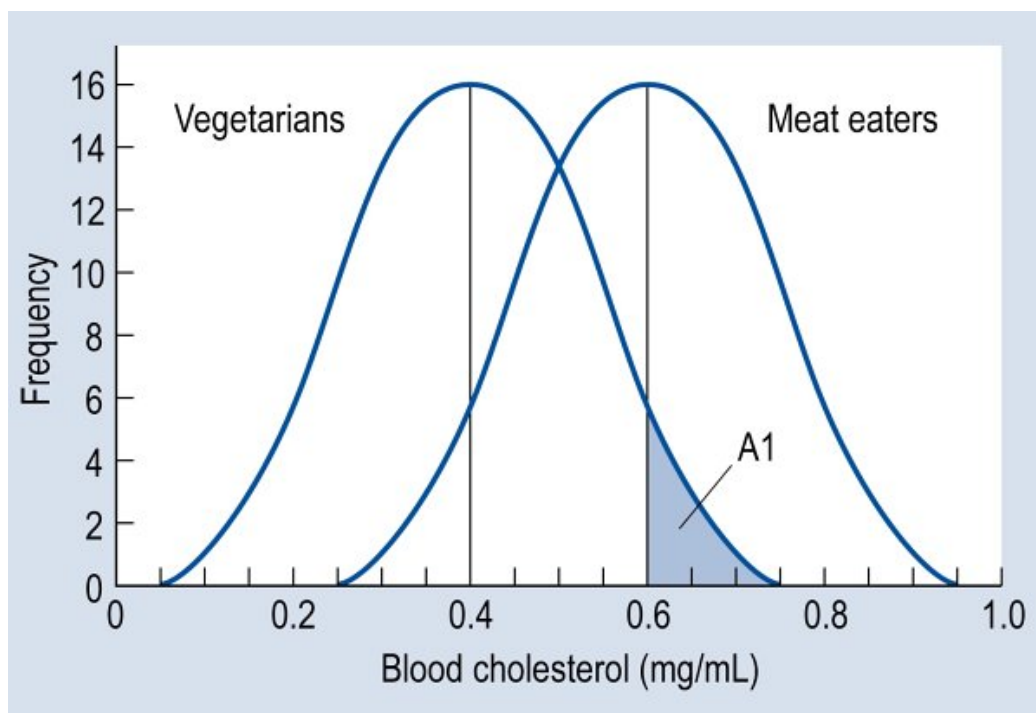


FIG. 16.6 Area A1 corresponds to the percentage of vegetarians with blood cholesterol higher than 0.6 mg/mL.

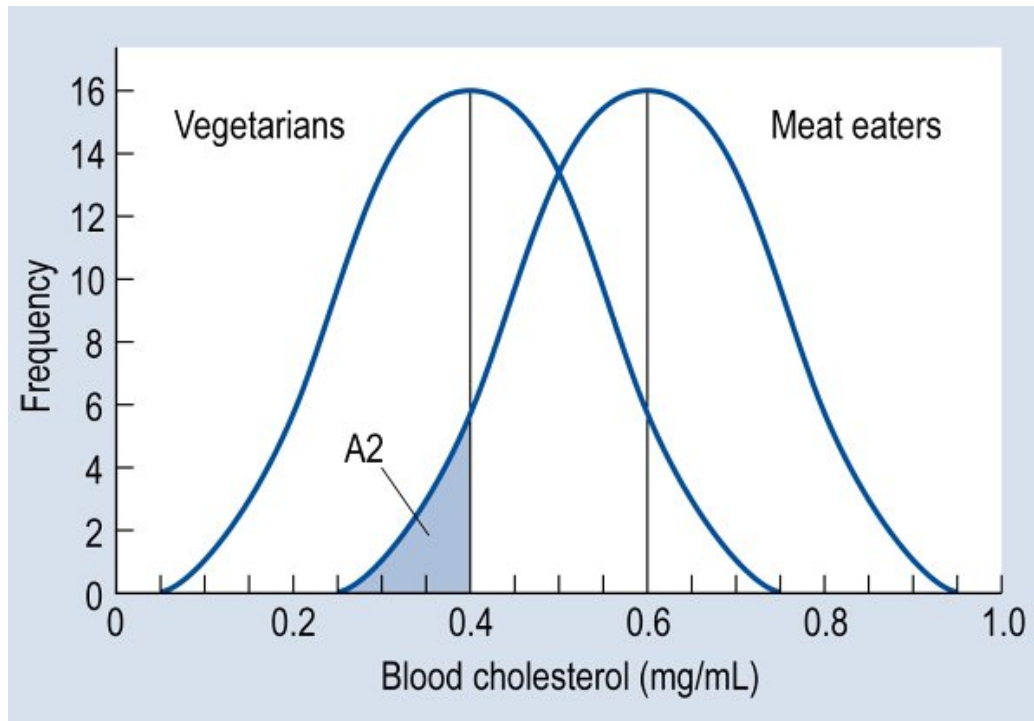


FIG. 16.7 Area A2 corresponds to the percentage of meat eaters with lower blood cholesterol than the mean for vegetarians.

Fig. 16.7 demonstrates the area (A2) corresponding to the percentage of meat eaters with lower blood cholesterol than the average vegetarian.

$$z = \frac{0.4 - 0.6}{0.15} = -1.33$$

Area A2 = 0.0918

Therefore approximately 9.2% of meat eaters had lower blood cholesterol levels than the average vegetarian.

Summary

We found that if the mean and standard deviation for a given distribution have been calculated, then we can transform any raw score into a standard (or z -) score. The z -score represents how many standard deviations a specific score is above or below the mean. The larger the value of the corresponding z -score (whether it is positive or negative), the more deviant

a given score is from the mean. A positive score means that it is above the mean and negative value means that it is below the mean. We described how to calculate this transformed score for a population or a sample. Also, we outlined the essential characteristics of the normal and the standard normal curves.

If the original frequency distribution is approximately normal, then the table of standardized normal distributions (Appendix A) can be used to calculate percentile ranks of raw scores, or the percentage of scores falling between specified scores. Also, z -scores are useful in comparing scores arising from two or more different normal distributions. The information presented here is applicable to clinical practice, for example in interpreting the significance of an individual's assessment in relation to known population norms. Many clinical test results routinely use standard score principles in their interpretation.

Correlation

Introduction,
Correlation,
Correlation Coefficients,
 Selection of Correlation Coefficients Suitable
 for Your Data,
 Calculation of Correlation Coefficients,
Uses of Correlations in the Health Sciences,
 Estimating Associations Between Variables,
 Reliability and Predictive Validity of
 Assessment,
 Estimating Shared Variance Using Correlation
 Coefficients,
Correlation and Causation,
Summary,

Introduction

A common aim of scientific and clinical research is to establish the nature of the relationships between two or more sets of observations or variables. Finding such relationships or associations can be an important step for identifying potential causal relationships and the prediction of clinical

outcomes. The topic of correlation is concerned with expressing quantitatively the size and the direction of the relationship between variables. Correlations are essential statistics in the health sciences, used to quantitatively measure the validity and reliability of clinical measures or expressing how health problems are associated with crucial biological, behavioural or environmental factors. After reading this chapter you will be able to explain how correlation coefficients are used and interpreted in health sciences research and practice.

The specific aims of this chapter are to:

1. Define the terms '*correlation*' and '*correlation coefficient*'.
2. Explain the correct selection and calculation of correlation coefficients.
3. Outline some of the research and clinical uses of correlation coefficients.
4. Define and calculate the coefficient of determination.
5. Discuss the relationship between correlation and causality.

Correlation

Consider the next two statements:

1. There is a positive relationship between cigarette smoking and lung damage.
2. There is a negative relationship between being overweight and life expectancy.

You probably have a fair idea what these two statements mean. The first statement implies that there is evidence that if you score high on one variable (cigarette smoking) you are likely to score high on the other variable (lung damage). The second statement describes the finding that scoring high on the variable 'overweight' tends to be associated with lowered 'life expectancy'. The information missing from each of the statements is the numerical value for size of the association between the variables. The variables in these statements could potentially be measured in a variety of different ways. For example, a common way to measure the

attribute overweight is to calculate body mass index (BMI; weight in kg/height in metres squared) but this method of measurement may work less well in some cultural groups (e.g. BMI is not a very good method of representing obesity among some Asian peoples).

A *correlation coefficient* is a statistic which expresses numerically the magnitude and direction of the association between two variables.

Table 17.1

Student	Score (out of 10)	
	Anatomy (X)	Physiology (Y)
1	3	2.5
2	4	3.5
3	1	0
4	8	6
5	2	1

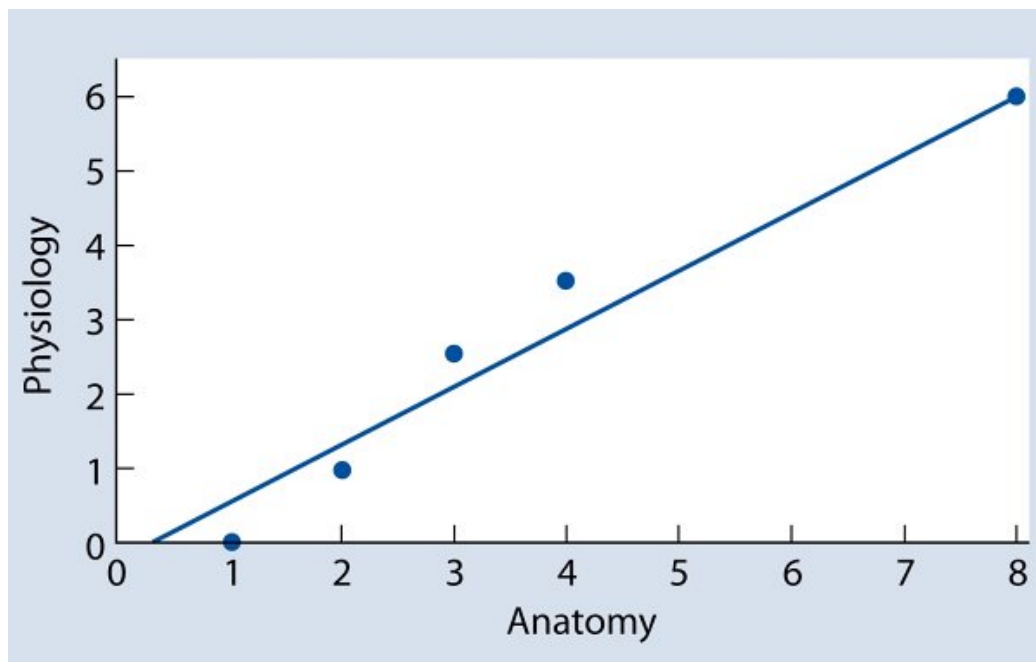


FIG. 17.1 Scattergram of students' scores in two examinations.

To demonstrate that two variables are correlated, we must obtain paired measures of both variables for the same set of participants or events. Let us consider an example to illustrate this point.

Assume that we are interested in determining whether student test scores for anatomy examinations are correlated with test scores for physiology. To keep the example simple, we will assume that there were only five ($n = 5$) students who sat for both examinations ([Table 17.1](#)).

To provide a visual representation of the relationship between the two variables, we can plot the above data on a *scattergram* (also referred to as a *scatterplot*). A scattergram is a graph of the paired scores for each participant on the two variables. By convention, we call one of the variables x and the other one y . It is evident from [Fig. 17.1](#) that there is a positive relationship or association between the two variables. That is, students who have high scores for anatomy (variable x) tend to have high scores for physiology (variable y). Also, for this set of data, we can fit a straight line in close approximation to the points on the scattergram. This line is referred to as a line of '*best fit*'. This topic is discussed further in statistics under '*linear regression*'. In general, a variety of relationships are possible between two variables (the scattergrams in [Fig. 17.2](#) illustrate some of these).

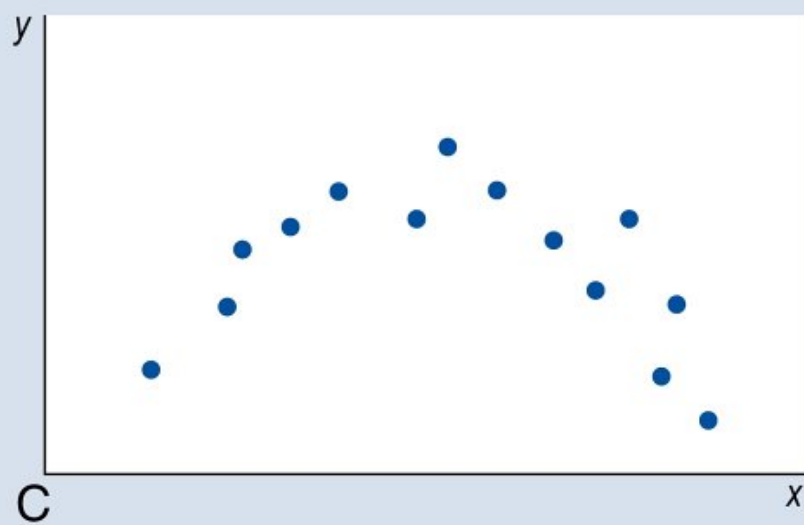
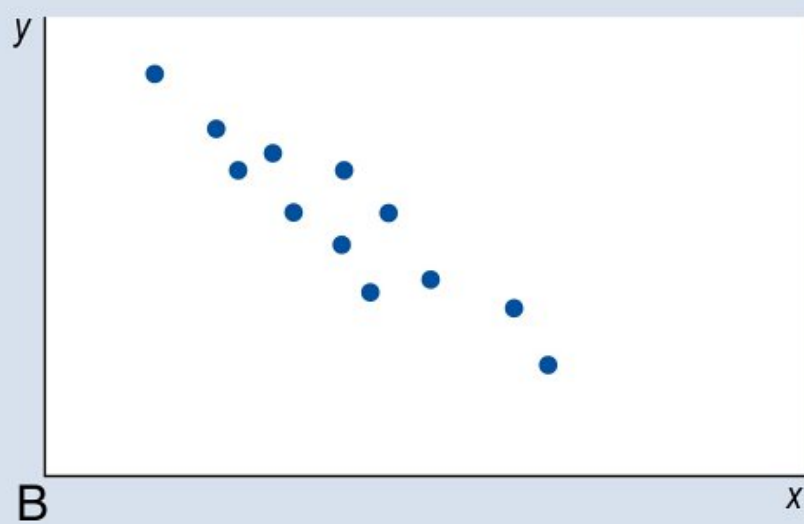
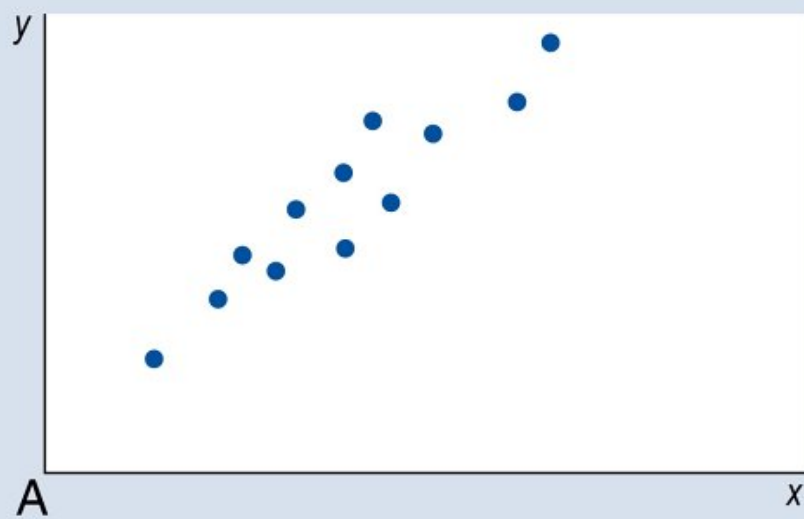


FIG. 17.2 Scattergrams showing relationships between two variables: (A) positive linear correlation; (B) negative linear correlation; (C) nonlinear correlation.

Fig. 17.2A and B represent a linear correlation between the variables x and y . That is, a straight line is the most appropriate representation of the relationship between x and y . Fig. 17.2C represents a nonlinear correlation, where a curve best represents the relationship between x and y .

Fig. 17.2A represents a *positive* correlation, indicating that high scores on x are related to high scores on y . For example, the relationship between cigarette smoking and lung damage is a positive correlation. Fig. 17.2B represents a *negative* correlation, where high scores on x are associated with low scores on y . For example, the correlation between the variables ‘being overweight’ and ‘life expectancy’ is negative, meaning that the more you are overweight, the lower your life expectancy.

Table 17.2

Correlation Coefficient

Coefficient	Conditions where appropriate
ϕ (phi)	Both x and y measures on a nominal scale
ρ (rho)	Both x and y measures on, or transformed to, ordinal scales
r	Both x and y measures on an interval or ratio scale

The graphical representation of the data pairs is an excellent way of performing an initial assessment of the nature of the association between two or more variables.

Correlation Coefficients

When we need to know or express the numerical value of the correlation between x and y , we calculate a statistic called the correlation coefficient. The correlation coefficient expresses quantitatively the magnitude and direction of the correlation between the two variables. Actually there is not a single “the” correlation coefficient, there are many correlation coefficients

(but with the same basic characteristics) that are targeted to calculating the associations between variables of different types. [Chapter 13](#) on measurement and observation provides a detailed discussion of these different data types.

Selection of Correlation Coefficients Suitable for Your Data

There are several types of correlation coefficients used in statistical analysis. [Table 17.2](#) shows some of these common correlation coefficients, and the conditions under which they are used. As [Table 17.2](#) indicates, the scale of measurements used determines the selection of the appropriate correlation coefficient.

All of the correlation coefficients shown in [Table 17.2](#) are appropriate for quantifying linear relationships between variables (or at least monotonic functions where the paired values only increase or decrease together). There are other correlation coefficients, such as η (eta) which are used for quantifying nonlinear relationships. Some relationships are very clear but are not linear. However, the discussion of the use and calculation of all the correlation coefficients is beyond the scope of this text. Rather, we will examine only the very commonly used Pearson's r and Spearman's ρ (rho). It is important to know that statistical packages such as SPSS, SAS provide simple means of performing these calculations. We have not focused on the detailed computation of lengthy examples because you are better off using a package to do this.

Regardless of which correlation coefficient we employ, these statistics share certain characteristics.

1. Correlation coefficients are calculated from pairs of measurements on variables x and y for the same group of individuals.
2. A positive correlation is denoted by $+$ and a negative correlation by $-$.
3. The values of the correlation coefficient range from $+1$ to -1 , where $+1$ means a perfect positive correlation, 0 means no correlation, and -1 a perfect negative correlation.

4. The square of the correlation coefficient represents the coefficient of determination.

Calculation of Correlation Coefficients

Pearson's r

We have already stated that Pearson's r is the appropriate correlation coefficient when both variables x and y are measured on an interval or a ratio scale. Further assumptions in using r are that both variables x and y are normally distributed, and that we are describing a linear (rather than a curvilinear) relationship.

Pearson's r is a measure of the extent to which paired scores are correlated. To calculate r , we need to represent the position of each paired score within its own distribution, so we convert each raw score to a z or standard score. Standard scores are calculated by converting the scores within a distribution or study sample to a distribution with a mean of 0 and a standard deviation of 1. This is easily achieved. This transformation corrects for the two distributions x and y having different means and standard deviations. The formula for calculating Pearson's r is:

$$r = \frac{\sum z_x z_y}{n}$$

where z_x = standard score corresponding to any raw x score, z_y = standard score corresponding to any raw y score, Σ = sum of standard score products and n = number of paired measurements.

[Table 17.3](#) gives the calculations for the correlation coefficient for the data given in the earlier examination scores example.

Therefore:

$$r = \frac{\sum z_x z_y}{n} = \frac{3.94}{5} = 0.79$$

(The z scores are calculated as discussed in [Ch. 16](#).)

Table 17.3

Student	Raw scores		z-scores		
	X	Y	z_x	z_y	$z_x z_y$
1	3	2.5	-0.22	-0.04	+0.01
2	4	3.5	+0.15	+0.39	+0.06
3	1	0	-0.96	-1.12	+1.08
4	8	6	+1.63	+1.46	+2.38
5	2	1	-0.59	-0.69	+0.41
	$\bar{x} = 3.6$	$\bar{y} = 2.6$			$\sum z_x z_y = 3.94$
	$s_x = 2.7$	$s_y = 2.33$			

This is quite a high correlation, indicating that the paired z -scores fall roughly on a straight line. In general, the closer the relationship between the two variables approximates a straight line, the more r approaches 1.0. Note that in social and biological sciences, correlations this high do not commonly occur. In general, we consider any correlation >0.7 to be quite high, 0.3 to 0.7 to be moderate and <0.3 to be weak. The scattergrams in [Fig. 17.3](#) illustrate this point.

When n is large, the previous equation is not sufficient for calculating r . Here we are not concerned with calculating r with a large n , although appropriate formulae or computer programs are available. For example, Pallant's survivor guide for SPSS describes how to use the program SPSS for calculating correlation coefficients.¹ The printouts not only provide the value of the correlation coefficients but also the statistical significance of the correlations

Assumptions for using Pearson's r

It was pointed out earlier that Pearson's r is used when the two variables are scaled on interval or ratio scales and when it is shown that they are linearly associated. Additionally, the sets of scores on each variable should be approximately normally distributed.

If any of these assumptions are violated, then the correlation coefficient might be spuriously low. Therefore other correlation coefficients should be used to represent association between two variables. A further problem may arise from the truncation of the range of values in one or both of the variables. This occurs when the distributions greatly deviate from normal shapes. The higher scale can be readily reduced to an ordinal scale.

If we measure the correlation between examination scores and intelligence quotients (IQs) of a group of health science students, we might find a low correlation because, by the time students present themselves to tertiary courses, most students with low IQs have been eliminated. In this way, the distribution of IQs would not be normal but rather negatively skewed. In effect, the question of appropriate sampling is also relevant to selecting correlation coefficients.

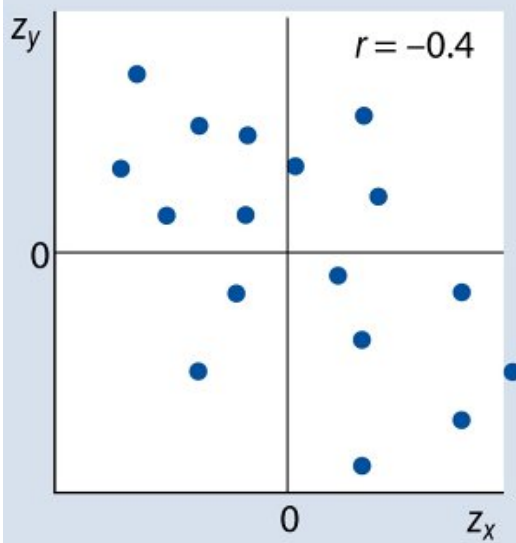
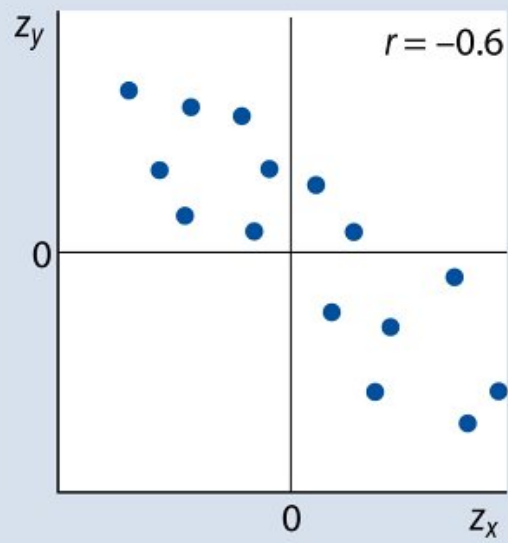
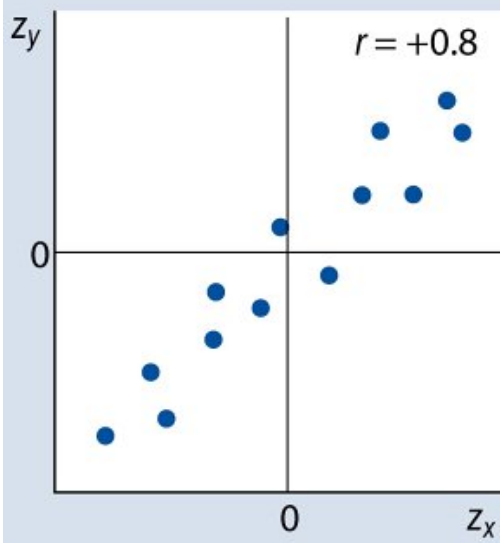
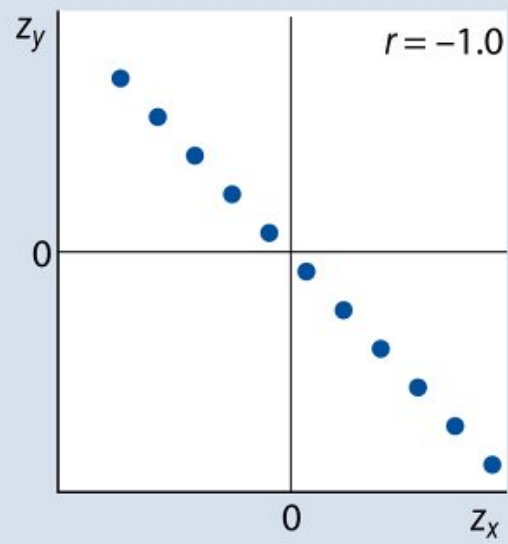
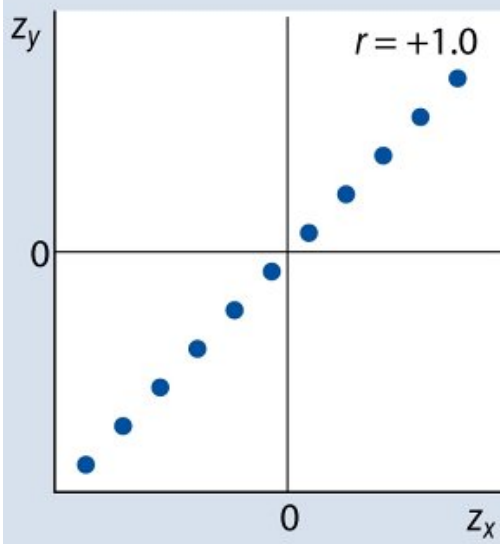


FIG. 17.3 Scattergrams and corresponding approximate r values.

Spearman's ρ

When the obtained data are such that at least one of the variables x or y was measured on an ordinal scale and the other on an ordinal scale or higher, we use ρ to calculate the correlation between the two variables. The higher scale can be readily reduced to an ordinal scale. If one or both variables were measured on a nominal scale, ρ is no longer appropriate as a statistic.

Table 17.4

Patient	Socio-economic status (Rank)	Severity of illness (Rank)	D (difference between ranks)	D^2
1	6	5	1	1
2	7	8	-1	1
3	2	4	-2	4
4	3	3	0	0
5	5	7	-2	4
6	4	1	3	9
7	1	2	-1	1
8	8	6	2	4
				$\Sigma d^2 = 24$

$$\rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

where d denotes the difference between paired ranks and n represents the number of pairs.

The derivation of this formula will not be discussed here. The 6 is placed in the formula as a scaling device; it ensures that the possible range rho is from -1 to $+1$ and thus enables ρ and r to be compared. Let us consider an example to illustrate the use of ρ .

If an investigator is interested in the correlation between socioeconomic status and severity of respiratory illness, and assuming that both variables were measured on, or transformed to, an ordinal scale, the investigator rank-orders the scores from highest to lowest on each variable ([Table 17.4](#)).

$$\text{Therefore : } \rho = 1 - \frac{6 \sum d^2}{n^3 - n}$$

$$1 - \frac{6 \times 24}{8^3 - 8} = 0.71$$

Clearly, the association among the ranks for the paired scores on the two variables becomes closer the more ρ approaches $+1$. If the ranks tend to be inverse, then ρ approaches -1 .

Uses of Correlations in the Health Sciences

Estimating Associations Between Variables

When the correlation coefficient has been calculated, it may be used to estimate the value of one variable (y) given the value of the other variable (x). For instance, take a hypothetical example that the correlation between cigarette smoking and lung damage is $r = +1$. We can see from [Fig. 17.4](#) that, given any score on x , we can transform this into a z score (z_x) and then using the graph, we can read off the corresponding z score on y (z_y).

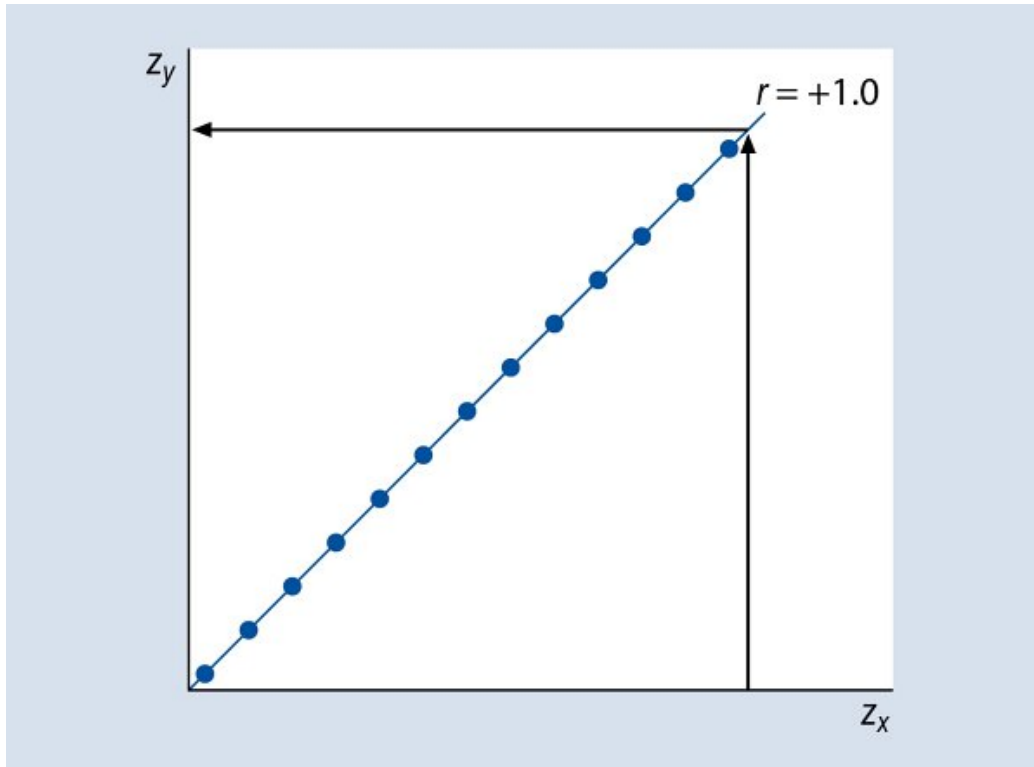


FIG. 17.4 Hypothetical relationship between cigarette smoking and lung damage. $r = +1$.

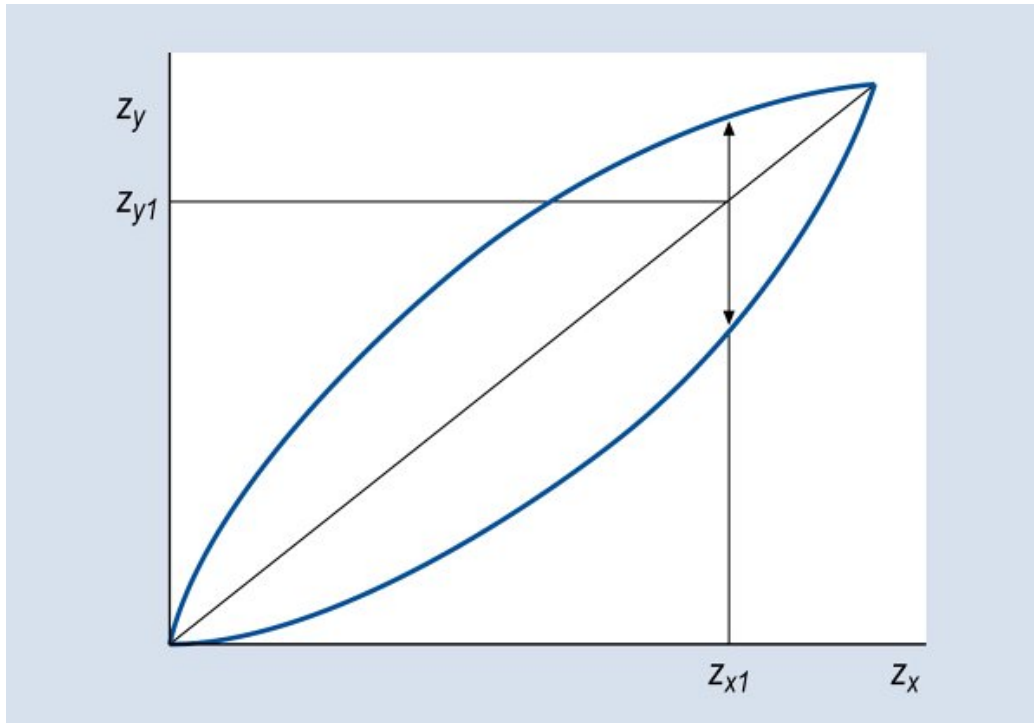


FIG. 17.5 Hypothetical relationship between cigarette smoking and lung damage. $r = +0.7$. With a correlation of <1.0 the data points will cluster around rather than exactly on the line and may vary between the range of values shown for any given values of x or y and their corresponding z -scores.

Of course, it is quite rare that there should be a perfect ($r = +1$) correlation between two naturally occurring variables. In this case the smaller the correlation coefficient, the greater the probability of making an error in estimation. For example, consider Fig. 17.5, where the scattergram shown represents a hypothetical correlation of approximately $r = 0.7$. Here, for any transformed value on variable x (say z_{x1}) there is a range of values of z_y that correspond. Our best guess here is z_{y1} , the average value, but clearly a range of scores is possible, as shown in Fig. 17.5. We will look at the issue of sampling error when we look at confidence intervals in Chapter 18. That is, as the correlation coefficient approaches 0, the range of error in prediction increases. A more appropriate and precise way of making estimations is in terms of regression analysis, but this topic is not covered in this introductory book.

Table 17.5

Patient	Observer X	Observer Y	Rank X	Rank Y
1	85%	98%	1	1
2	75%	86%	2	2
3	74%	80%	3	3
4	72%	79%	4	4
5	69%	70%	5	5

Reliability and Predictive Validity of Assessment

As you will recall, reliability refers to measurements using instruments or to subjective judgements remaining relatively the same on repeated administration. This is called test–retest reliability and its degree is measured by a correlation coefficient. The correlation coefficient can also be used to determine the degree of inter-observer reliability.

As an example, assume that we are interested in the inter-observer reliability (agreement) for two neurologists who are assessing patients for degrees of functional disability after spinal cord damage. [Table 17.5](#) represents a set of hypothetical results of their independent assessment of the same set of five patients. There are many studies of inter-observer agreement in clinical judgements in the medical literature because such judgements are often very influential in decisions about treatment and diagnosis.

Observer Y clearly attributes greater degrees of disability to the patients than observer X. However, as stated earlier, this need not affect the correlation coefficient. If we treat the measurement as an ordinal scale, we can see from [Table 17.5](#) that the ranks given to the patients by the two observers correspond, so that it can be shown that $\rho = +1$.

Clearly, the higher the correlation the greater the reliability. If we had treated the measurements as representing interval or ratio data, we would have calculated Pearson's r to represent quantitatively the reliability of the measurement.

Predictive validity is also expressed as a correlation coefficient. For example, say that you devise an assessment procedure to predict how much

people will benefit from a rehabilitation program. If the correlation between the assessment and a measure of the success of the rehabilitation program is high (i.e. 0.7 or 0.8), the assessment procedure has high predictive validity. If, however, the correlation is low (i.e. 0.3 or 0.4), the predictive validity of the assessment is low, and it would be unwise to use the assessment as a screening procedure for the entry of patients into the rehabilitation program.

Estimating Shared Variance Using Correlation Coefficients

A useful statistic is the square of the correlation coefficient (r^2), which represents the proportion of variance in one variable accounted for by the other. This is called the *coefficient of determination*.

If, for example, the correlation between variable x (height) and variable y (weight) is $r = 0.7$, then the coefficient of determination is $r^2 = 0.49$, or 49%. This means that 49% of the variability of weight can be accounted for in terms of height. You might ask, what about the other 51% of the variability? This would be accounted for by other factors, say for instance, a tendency to eat fatty foods. The point here is that even a relatively high correlation coefficient ($r = 0.7$) accounts for <50% of the variability.

This is a difficult concept, so it might be worth remembering that ‘variability’ (see [Ch. 15](#)) refers to how scores are spread out about the mean. That is, as in the example just given, some people will be heavy, some average and some light. So we can account for 49% of the total variability of weight (x) in terms of height (y) if $r = 0.7$. The other 51% can be explained in terms of other factors, such as somatotype. The greater the correlation coefficient, the greater the coefficient of determination, and the more the variability in y can be accounted for in terms of x .

Correlation and Causation

Previously, we pointed out that there were at least three criteria for establishing a causal relationship. Correlation or covariation is only one of these criteria. We have already discussed that even a high correlation between two variables does not necessarily imply a causal relationship.

That is, there are a variety of plausible explanations for finding a correlation. As an example, let us take cigarette smoking (x) and lung damage (y). A high positive correlation could result from any of the circumstances presented here:

1. x causes y .
2. y causes x .
3. There is a third variable, which causes changes in both x and y .
4. The correlation represents a spurious or chance association.

For this example, (1) would imply that cigarette smoking causes lung damage, (2) that lung damage causes cigarette smoking, and (3) that there was a variable (e.g. stress) which caused both increased smoking and lung damage. We need further information to identify which of the competing hypotheses is most plausible.

Some associations between variables are completely spurious (4). For example, there might be a correlation between the amount of margarine consumed and the number of cases of influenza over a period of time in a community, but each of the two events might have entirely different, unrelated causes. Also, some correlation coefficients do not reach statistical significance: that is, they may be due to chance.

If we are using a sample to estimate the correlation in a population, we must be certain that the outcome is not due to biased sampling or sampling error. That is, the investigator needs to show that a correlation coefficient is statistically significant and not just due to random sampling error. We will look at the concept of statistical significance in [Chapter 18](#).

As we demonstrate that correlation does not establish causality, we can use correlations as a source for subsequent hypotheses. For instance, in this example, work on carcinogens in cigarette tars and the causes of emphysema have confirmed that it is probably true that smoking does, in fact, cause lung damage (x causes y). However, given that there are often multiple causations of health problems, option (3) cannot be ruled out.

Techniques are available which can, under favourable conditions, enable investigators to use correlations as a basis for distinguishing between competing plausible hypotheses. One such technique, called *path analysis*, involves postulating a probable temporal sequence of events and then

establishing a correlation between them. This technique was borrowed from systems theory and has been applied in areas of clinical research, such as epidemiology.

Summary

In this chapter, we outlined how the degree and direction between two variables can be quantified using statistics called correlation coefficients. Of the correlation coefficients, the use of two was outlined: Pearson's r and Spearman's ρ . Definitional formulae and simple calculations were presented, with the understanding that more complex calculations of correlation coefficients are done with computers.

Several uses for r and ρ were discussed for prediction, for quantifying the reliability and validity of measurements, and by using r^2 in estimating the amount of shared variability. Finally, we discussed the caution necessary in causal interpretation of correlation coefficients. Showing a strong association between variables is an important step towards establishing causal links but further evidence is required to decide the direction of causality.

As with other descriptive statistics, caution is necessary when correlation coefficients are calculated for a sample and then generalized to a population. Generalization from a sample statistic to a population parameter must take into account the probability of sampling error, an issue which will be discussed in the following chapters.

Reference

1. Pallant J. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*. Melbourne: Allen and Unwin; 2016.

SECTION 6

Data Analysis and Inference

OUTLINE

Introduction

18. Probability and Confidence Intervals

19. Hypothesis Testing

20. Effect Size and the Interpretation of Evidence

21. Qualitative Data Analysis

22. Critical Evaluation of Published Research

Introduction

Statistical data analysis is an essential activity in conducting quantitative research (see [Ch. 3](#)). Quantitative research in the health sciences mostly involves working with smaller clinical samples drawn from larger clinical populations. To generalize our findings, we must draw generalizations and inferences from sample statistics which describe the characteristics of a sample (e.g. $\bar{x}$, s) to the population parameters which describe the characteristics of the target population (e.g. μ , σ). There is always uncertainty in inferential statistics because when samples are drawn, their characteristics may differ from the population characteristics due to sampling error (the discrepancy between the sample statistics and the population parameters). The results obtained from our sample data could represent random variations or chance patterns rather than ‘real’ ones which are true for the population as a whole. If the sampling method was biased, then the differences could be quite large. Similarly, if the study sample is quite small then there may be significant variability in the obtained results from one sample to another.

As an illustration, imagine that we have collected data in a study aimed at identifying age-related differences in the use of sedatives and tranquillizers in a given population. Our participants ($n = 200$) kept diaries over a period of 1 year, recording each time they had taken a sedative or tranquillizer. The research question is: Is there a difference between sedative or tranquillizer use in older and younger people? Assume that the hypothetical data presented here were obtained:

Sedative and tranquillizer consumption			
Age group	n		s
20–39	100	20	5
40–59	100	30	5

Three important and interrelated questions are examined in Section 6 concerning the evidence provided by sample data such as those shown in the table.

1. Even if we used an adequate sampling procedure (see [Ch. 5](#)), how confident are we to infer that the true population parameters (μ) are the same as, or are at least close to, ? For example, is it true that the mean tranquillizer intake for the 20 to 39 age group is $\mu = 20$?
2. It appears that there is a large difference between the two sample means; but is this difference also true for the populations? In other words, is this difference 'real' or significant, or is it simply due to sampling error?
3. What we are inferring is that $\mu(\text{older}) > \mu(\text{younger})$. In other words, are we justified in concluding that the mean sedative/tranquillizer intake for the older age group is greater than for the younger group in the population?

The key issue here is that we are using sample data for decisions about populations. As previously stated, the sampling error is the difference between the obtained sample statistics and the actual state of the population. We cannot eliminate sampling error even with large and well-chosen samples. Rather, we can apply the principles of inferential statistics to calculate the probability of error. We then use this information to minimize the probability of making errors when we generalize from sample statistics to population parameters.

In [Chapter 19](#), we examine how sampling distributions are derived and used for calculating the probability of obtaining a given sample statistic. This information can be applied to the calculation of confidence intervals, which represent a range of scores which contain the true population parameter at a given level or probability.

In [Chapter 20](#), we outline the logic of hypothesis testing, using single sample z and t tests as exemplars. Hypothesis testing is a procedure used to decide if a difference or pattern identified in our sample data is statistically significant. If the outcome of our analysis is significant, then we are in a position to conclude that the patterns or differences found in our data may be generalized to the populations from which the samples were drawn.

There are numerous statistical tests available for analysing the significance of our data. In [Chapter 20](#), we discuss criteria for selecting an appropriate statistical test, including (1) scale of measurement used to collect the data, (2) the number of groups being compared and (3) the dependence or independence of measurements. We will use the χ^2 test to demonstrate how statistical tests are selected and used to analyse the data.

Ultimately, statistical decision making is probabilistic (i.e. has some degree of chance and uncertainty), implying that the possibility of making decision errors cannot be eliminated. Statistical decisions may be correct, or involve what are called type I and type II errors. In [Chapter 21](#), we examine how these errors may influence our interpretation of the results in relation to the aims or hypotheses guiding our research, and we will examine how to reduce the probability of making such errors.

The key concept discussed in [Chapter 21](#) is effect size. Effect size determines the utility (i.e. the clinical significance), of the research findings. Also in this chapter, we outline the relationship between effect size and the clinical or practical significance of research findings.

The aim of qualitative data analysis is to interpret the meaning of the experiences reported by the research participants. In [Chapter 22](#), we will discuss the basic principles of qualitative data analysis including coding, content analysis and thematic analysis.

Probability and Confidence Intervals

Introduction,
Probability,
Sampling Distributions,
Sampling Distribution of the Mean,
Calculating Confidence Intervals,
Confidence Intervals Using the t Distribution,
Confidence Intervals in Health Research,
Summary,

Introduction

Sample statistics (such as $\bar{x}$, s) are estimates of the actual population parameters, in this case μ , σ . Even where adequate sampling procedures are adopted, there is no guarantee that the sample statistics are exactly the same as the true parameters of the population from which the samples were drawn because of random sampling variation. Therefore inferences from sample statistics to population parameters necessarily involve the possibility of sampling error. Given that the researchers usually have no

knowledge of the true population parameters (because they are unable to study the entire population), inferential statistics are employed to estimate the probable sampling errors when using statistical data based on a study sample. Although sampling error cannot be completely eliminated, the probable size of sampling error can be calculated using *inferential statistics*. In this way investigators are in a position to calculate the probability of being accurate in their estimations of the actual population parameters.

The aims of this chapter are to examine how probability theory is applied to generating sampling distributions and how sampling distributions are used for estimating population parameters. Sampling distributions are used to estimate sampling error. This statistic is then used to calculate confidence intervals and testing hypotheses (see [Chapters 18](#) and [19](#)).

The specific aims of this chapter are to:

1. Define probability.
2. Demonstrate how sampling distributions are generated.
3. Show how sampling distributions of the mean are used to calculate the probability of a sample mean.
4. Explain how confidence intervals are calculated for continuous data.
5. Distinguish between z and t distributions.

Probability

The concept of probability is central to the understanding of inferential statistics. Probability is expressed as a proportion between 0 and 1, where 0 means an event is certain not to occur, and 1 means an event is certain to occur. Therefore if the probability (p) is 0.01 for an event then it is unlikely to occur frequently (the actual chance is 1 in 100). If $p = 0.99$ then the event is highly likely to occur (the chance is 99 times in 100). The probability of any event (say event A) occurring is given by the formula:

$$p(A) = \frac{\text{number of occurrences of A}}{\text{total number of possible occurrences}}$$

Sometimes the probability of an event can be calculated a priori (before the event) by logical reasoning. For example, we can predict that the probability of throwing a head (H) with a fair coin is:

$$p(H) = \frac{\text{number of occurrences of H}}{\text{total number of possible occurrences}}$$

$$\frac{H}{H + T \text{ (tails)}} = \frac{1}{2} = 0.5$$

Table 18.1

Causes of Death for Persons over 65

Cause of death	Percentage of deaths
Coronary heart disease	50
Cancer	25
Stroke	10
Accident	5
Infection	5
Other causes	5

Or, if we buy a lottery ticket in a (non-rigged!) lottery draw where there are 100 000 tickets, the probability of winning first prize is:

$$p(1 \text{ st prize}) = \frac{1}{100\,000} = 0.00001$$

In some situations, there are no theoretical grounds for calculating the occurrence of a single event (a priori). For instance, how can we calculate

the probability of a specific individual dying of a specific condition? Thus for example the probability of dying from infective endocarditis might be 0.25 in the first year after infection but for a specific individual we do not know whether they will die or not in the first year. In such instances, we use previously obtained evidence to calculate probabilities a posteriori (after the event).

For example, if we have information for the mortality rates of a community ([Table 18.1](#)) we are in a position to calculate the probability of a selected individual over 65 years of age dying of any of the specified causes. For example, the probability of a given individual dying of coronary heart disease is:

$$\begin{array}{l} p(\text{dying of heart disease}) \\ \frac{\% \text{ of cases dying of heart disease}}{100 \%} \\ \frac{5}{100} = 0.05 \end{array}$$

Also, we can use the normal curve model, as outlined in [Chapter 18](#), to determine the proportion or percentage of cases up to, or between, any specified scores. In this instance, probability is defined as the proportion of the total area cut off by the specified scores under the normal curve. As we discussed in [Chapter 17](#), the greater the area under the curve, the higher the corresponding probability of selecting specified values.

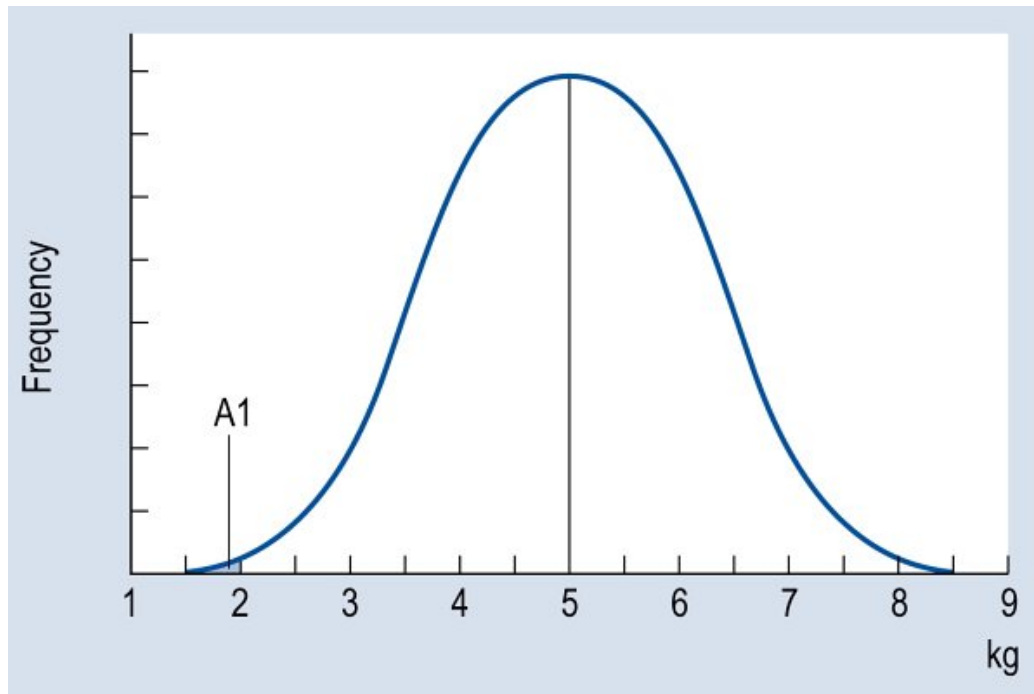


FIG. 18.1 Frequency distribution of neonate birth weights. Area A1 corresponds to $z < -2$.

For example, say that [Fig. 18.1](#) illustrates the birthweights of a large sample of neonates. Let us assume that the distribution is approximately normal, with the mean ($\bar{x}$) of 5 kg and a standard deviation (s) of 1.5. We can use the information to calculate the probability of any range of birth weights. Now, say that we are interested in the probability of a randomly selected neonate having a birth weight of 2 kg or under.

The area A1 under the curve in [Fig. 18.1](#) corresponds to the probability of obtaining a score of 2 or under. To calculate proportions or areas under the normal curve, we first translate the raw score of 2 into a z score:

$$z = \frac{x - \bar{X}}{s} = \frac{2 - 5}{1.5} = -2$$

Now we look up the area under the normal curve corresponding to $z = -2$ (Appendix A). Here we find that A1 is 0.0228. This area corresponds to a probability, and we can say that ‘The probability of a neonate having a

birthweight of 2 kg or less is 0.0228'. Another way of stating this outcome is that the chances are approximately 2 in 100, or 2%, for a child having such a birthweight.

We can also use the normal curve model to calculate the probability of selecting scores between any given values of a normally distributed continuous variable. For example, if we are interested in the probability of birthweights being between 6 and 8 kg, then this can be represented on the normal curve (area A2 on [Fig. 18.2](#)). To determine this area, we proceed as shown. Let $s = 1.5$.

$$z_1 = \frac{6-5}{1.5} = 0.67$$
$$z_2 = \frac{8-5}{1.5} = 2.0$$

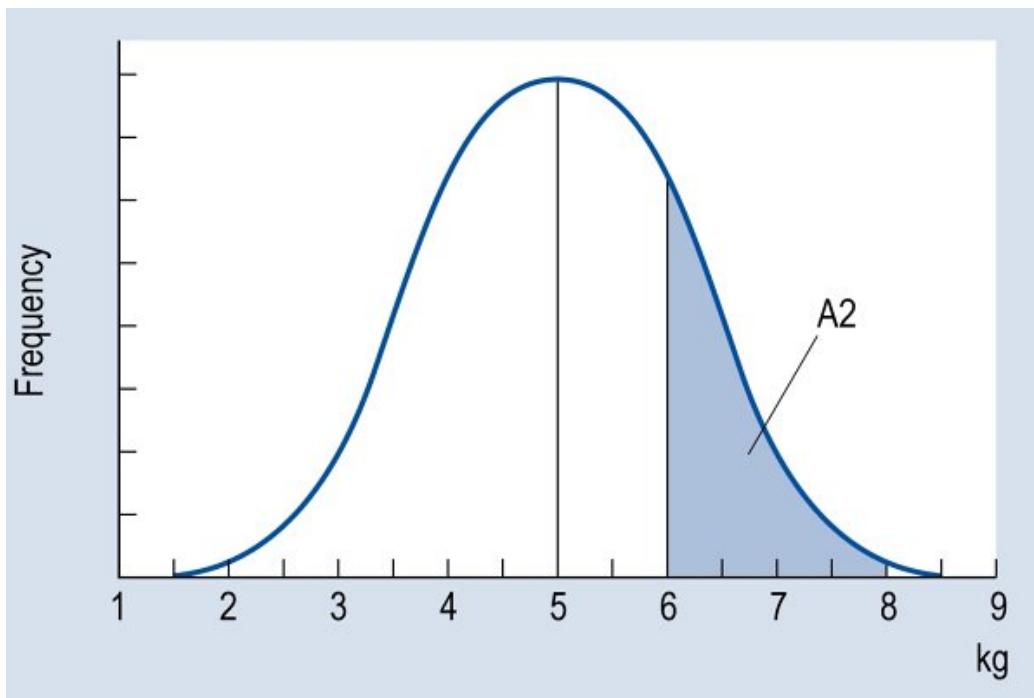


FIG. 18.2 Frequency distribution of neonate birthweights. Area A2 corresponds to probability of weight being 6 to 8 kg.

Therefore the area between z_1 and $\bar{x}$ is 0.2486 (from Appendix A) and the area between z_2 and $\bar{x}$ is 0.4772 (from Appendix A). Therefore the required area A2 is:

$$A2 = 0.4772 - 0.2486 = 0.2286$$

It can be concluded that the probability of a randomly selected child having a birthweight between 6 and 8 kg is $p = 0.2286$. Another way of saying this is that there is a 23 in 100, or a 23%, chance that the birthweight will be between 6 and 8 kg.

These examples demonstrate that, when the mean and standard deviation are known for a normally distributed continuous variable, this information can be applied to calculating the probability of events related to this distribution. Of course, probabilities can be calculated for other than normal data but this requires integral calculus, which is beyond the scope of this text. In general, regardless of the shape or scaling of a distribution, scores which are common or close to the average are more likely to be selected than those which are atypical, located at the extreme ends of a distribution.

Sampling Distributions

Probability theory also can be applied to calculate the likelihood of obtaining specific sample statistics from a specified population. These calculations are the basis for making probabilistic statistical inferences. Let us look at a hypothetical example of drawing samples for dichotomous outcomes.

Consider a container with a very large number of identically sized marbles. Imagine that there are two kinds of marbles present, black (B) and white (W), and that these colours are present in equal proportions, so that $p(B) = p(W) = 0.5$.

Possible outcomes	Number black	Number white	Proportions black
● ● ● ●	4	0	1.00
● ● ● ○	3	1	0.75
● ● ○ ○	2	2	0.50
● ○ ○ ○	1	3	0.25
○ ○ ○ ○	0	4	0.00

FIG. 18.3 Characteristics of possible samples of $n = 4$, drawn from a population of black and white marbles.

Given the above population, say that samples of marbles are drawn randomly and with replacement. (By ‘replacement’ we mean the samples are put back into the population to maintain as a constant the proportion of $B = W = 0.5$.) If we draw samples of four (i.e. $n = 4$) then the possible proportions of black and white marbles in the samples can be deduced a priori as shown in [Fig. 18.3](#).

Ignoring the order in which marbles are chosen, [Fig. 18.3](#) demonstrates all the possible outcomes for samples of sample size $n = 4$. It is logically possible to draw any of the samples shown. However, only one of the samples (2B, 2W) is representative of the true population parameter. The other samples would generate incorrect inferences concerning the state of the population. In general, if we know or assume (hypothesize) the true population parameters, we can generate distributions of the probability of obtaining samples of a given characteristic.

In this instance, when attempting to predict the probability of specific samples drawn from a population with two discrete elements, the binomial theorem can be applied. The expansion of the binomial expression $(P + Q)^n$ generates the probability of all the possible samples which can be drawn from a given population. The general equation for expanding the binomial expression is:

$$(P + Q)^n = P^n + \frac{n}{1} P^{n-1} Q + \frac{n(n-1)}{2} P^{n-2} Q^2 + \dots + Q^n$$

P is the probability of the first outcome, Q is the probability of the second outcome and n is the number of trials (or the sample size).

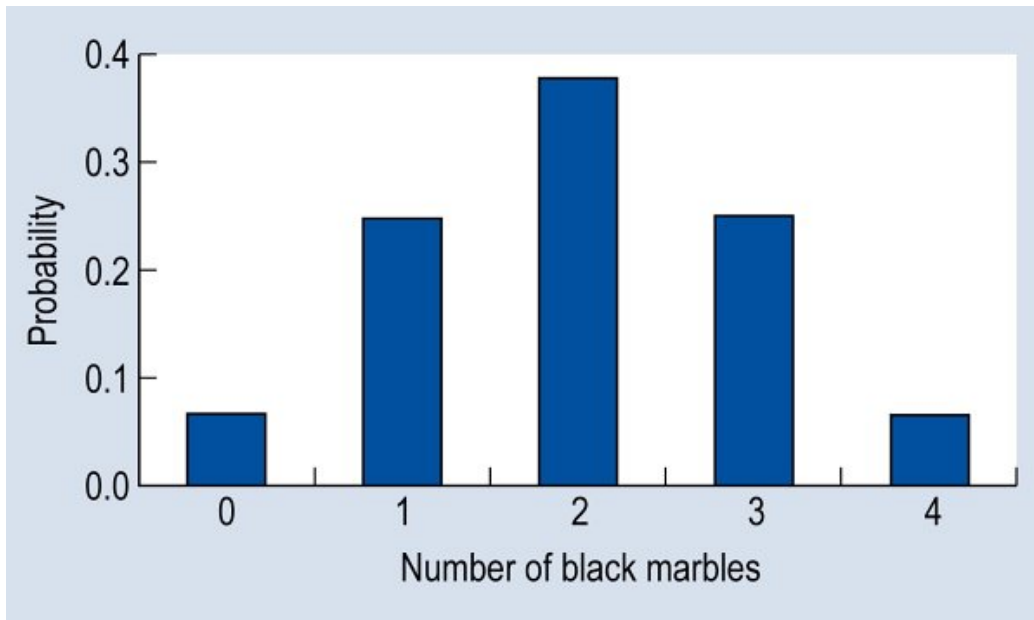


FIG. 18.4 Sampling distribution of black marbles, $n = 4$.

In this instance, $P = \text{proportion black (B)} = 0.5$, $Q = \text{proportion white (W)} = 0.5$, $n = 4$ (sample size). Therefore substituting into the binomial expression:

$$(B + W)^4 = B^4 + 4B^3W + 6B^2W^2 + 4BW^3 + W^4$$

The following shows the composition of the samples, which can be drawn from the specified population and the probability of obtaining a specific sample. For the present case:

$$\begin{aligned}
\text{Sample 1 } p(4B0W) &= B^4 = (0.5)^4 = 0.0625 \\
\text{Sample 2 } p(3B1W) &= 4B^3W = 4 \times (0.5)^3 (0.5) = 0.2500 \\
\text{Sample 3 } p(2B2W) &= 6B^2W^2 = 6 \times (0.5)^2 (0.5)^2 = 0.3750 \\
\text{Sample 4 } p(1B3W) &= 4BW^3 = 4 \times (0.5)^3 (0.5) = 0.2500 \\
\text{Sample 5 } p(0B4W) &= W^4 = (0.5)^4 = 0.0625
\end{aligned}$$

The calculated probabilities add up to 1, indicating that all the possible sample outcomes have been accounted for (because the probability of all possible events must equal 1). However, the important issue here is not so much the mathematical details but the general principle being illustrated by the example. For a given sample size (n) we can employ a mathematical formula to calculate the probability of obtaining all the possible samples from a population with known characteristics. The relationship between the possible samples and their probabilities can be graphed, as shown in [Fig. 18.4](#).

Taking the statistic ‘number of black marbles in the sample’, the graph in [Fig. 18.4](#) shows the probability of obtaining any of the outcomes. The distribution shown is called a ‘sampling distribution’. In general, a sampling distribution for a statistic indicates the probability of obtaining any of the possible values of a statistic.

Therefore having obtained our sampling distribution we can see that some sample outcomes have low probability, whereas others are more likely to occur. Although there is a finite chance of obtaining a sample such as ‘all blacks’, the probability of this happening is rather small ($p = 0.0625$). Conversely, a sample of 2B2W, which is equal to the true population proportions, is far more probable ($p = 0.375$).

This example might be more meaningful in a health research context if we say that the black marbles represent unfavourable treatment outcomes and the white marbles successful treatment outcomes. What the example shows is that, even if there were no actual differences in the outcomes, in some extreme samples it might seem that all the patients benefited (5W) or that none had (5B). What was demonstrated is that the probability for such extreme outcomes was improbable ($p = 0.0625$). Generating sampling distributions for calculating the probability of given sample statistics is a basic practice in inferential statistics. The sampling distributions enable

researchers to infer (with a determined level of confidence) the true population parameters from the sample statistics.

Sampling Distribution of the Mean

The binomial theorem is appropriate for generating sampling distributions for discontinuous nominal scale data. When measurements are continuous, the mean and standard deviations are appropriate because sample statistics are measured on interval or ratio scales. The sampling distribution of the mean represents the frequency distribution of sample means obtained from random samples drawn from the population. The sampling distribution of the mean enables the calculation of the probability of obtaining any given sample mean ($\bar{x}$). This is essential for testing hypotheses about sample means (see [Ch. 19](#)).

To generate the sampling distribution of the mean, we use a mathematical theorem called the central limit theorem. This theorem provides a set of rules which relate the parameters (μ , σ) of the population from which samples are drawn to the distribution of sample means ($\bar{x}$).

The central limit theorem states that, if random samples of a fixed n are drawn from any population, as n becomes large the distribution of sample means approaches a normal distribution, with the mean of the sample means ($\bar{\bar{x}}$ or $\mu_{\bar{x}}$) being equal to the population mean (μ) and the standard error of estimate ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$) being equal to $\sigma/\sqrt{n}$. The standard error of the estimate is the standard deviation of the distribution of sample means.

Let us follow this step by step.

1. Imagine we have a population of continuous scores or measurements with a mean of μ and a standard deviation of σ .
2. We select a very large number of random samples, each sample being of a size n .
3. Having obtained our samples, for each sample we calculate the sample mean ($\bar{x}_1, \bar{x}_2, \dots$ and so on).
4. Each sample mean, $\bar{x}$, is a number. The sampling distribution of the mean is a frequency distribution representing the large number of sample means.

5. The central limit theorem predicts theoretically the shape (normal for large n), mean ($\bar{\bar{x}}$ or $\mu_{\bar{x}}$) and standard deviation $s_{\bar{x}}$ or $\sigma_{\bar{x}}$ of a large number of sample means.

It should be noted that:

1. The sampling distribution of the mean is a frequency distribution of a large number of sample means of size n drawn from a given population. When n increases, the sampling distributions approach normal.
2. The mean of the sample means $\mu_{\bar{x}}$ or $\bar{\bar{x}}$ is the mean of the distribution of sample means. $\bar{\bar{x}}$ and $\mu_{\bar{x}}$ are equal to μ , the population mean.
3. The standard error of the mean $s_{\bar{x}}$ or $\sigma_{\bar{x}}$ is the standard deviation of the frequency distribution of sample means drawn from a population. The magnitude of $s_{\bar{x}}$ or $\sigma_{\bar{x}}$ is equal to $\sigma/\sqrt{n}$, the population standard deviation divided by the square root of the sample size.
4. $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ are used in reference to a sampling distribution based on all the possible samples drawn from a population (a population of samples, would you believe?), whereas $\bar{\bar{x}}$ and $s_{\bar{x}}$ are used when the sampling distribution is based on a 'sample' of samples.

Let us have a look at an example. Assume for a hypothetical test of motor function that $\mu = 50$ and $\sigma = 10$. What is the probability of drawing a random sample from this population with $\bar{x} = 52$ or greater (i.e. $\bar{x} > 52$) given that $n = 100$? The central limit theorem predicts that when we draw samples of $n = 100$ from the above population, the sampling distribution of the means will be as follows:

- The shape of the sampling distribution will be approximately normal.
- The mean of the sampling distribution will be equal to μ :

$$\mu_{\bar{x}} = \mu = 50$$

- The standard error of estimate ($\sigma_{\bar{x}}$) will be:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

(We can show this as in [Fig. 18.5](#).)

Previously, we saw how we can use normal frequency distributions for estimating probabilities. Using the same principles as in [Chapter 17](#), we can calculate the z score corresponding to $\bar{x} = 52$, and look up Appendix A to find out the area representing the probability in question:

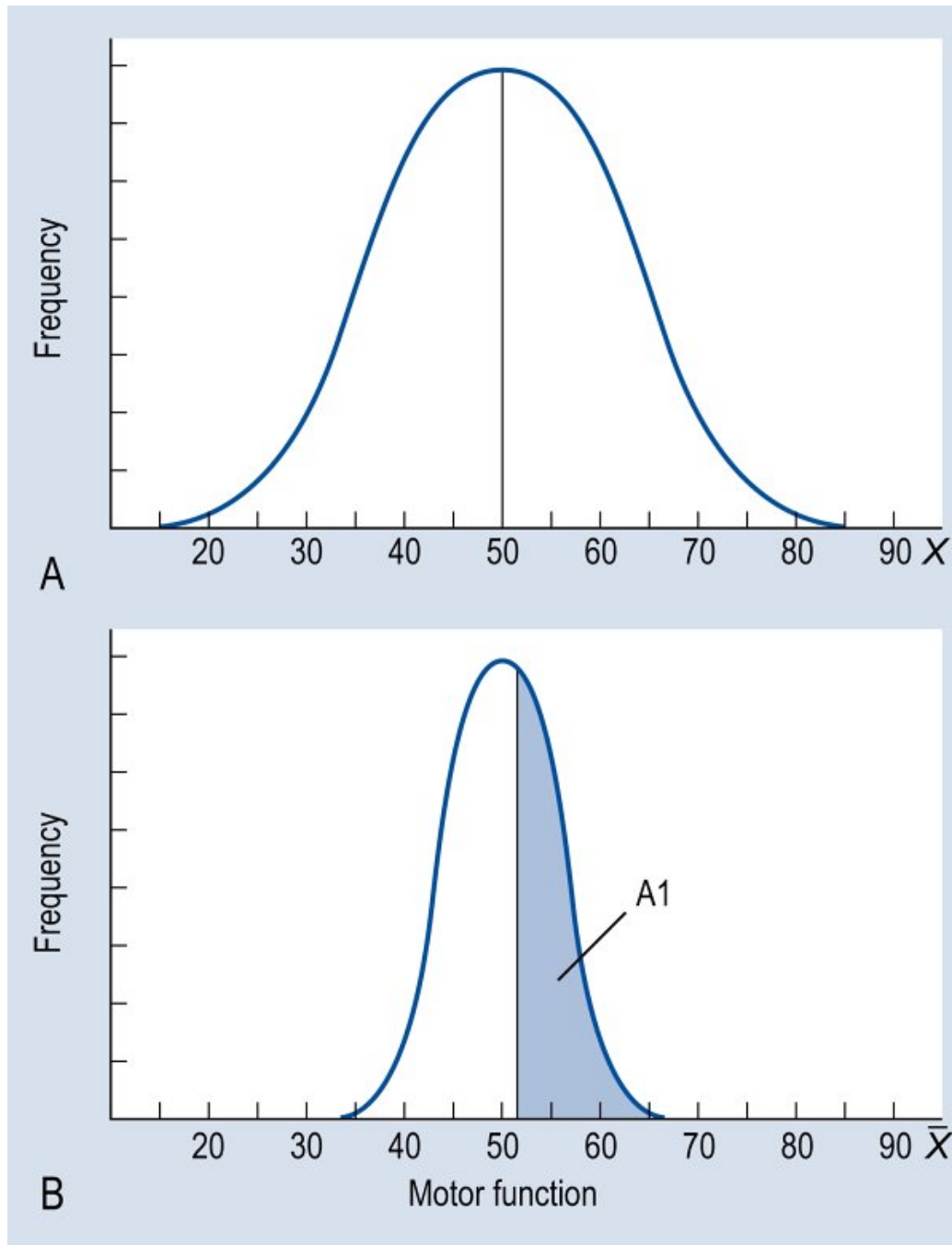


FIG. 18.5 Relationship between original population and sampling distribution of the mean, where $\mu = 50$ and $\sigma = 10$: (A) original population; (B) sampling distribution of the mean.

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{52 - 50}{1} = 2$$

That is, $\bar{x} = 52$ is two standard error units above μ_x , the population mean for the sample means.

You may have noticed that the distribution of $\bar{x}$ is far less dispersed than X (i.e. the raw scores), as $\sigma_{\bar{x}} = 1$ and $\sigma = 10$.

Using Appendix A for establishing the probability, we find that the area representing $p(\bar{x} > 52)$ (i.e. area A1 in [Fig. 18.5B](#)) is:

$$p(\bar{X} > 52) = 0.5000 - 0.4772 = 0.0228$$

Therefore the probability of drawing a sample of $\bar{x} > 52$ is 0.0228. We will apply this notion to hypothesis testing in [Chapter 19](#), but you might have noticed that it is a rather low probability. That is, it is unlikely ($p = 0.0228$) randomly to draw a sample with $\bar{x} > 52$ where $n = 100$ and $\mu = 50$.

Calculating Confidence Intervals

Let us assume that you are asked to estimate the weight of a newborn baby. If you have experienced working with neonates, you should be able to make a reasonable guess. You might say ‘The baby is 6 kg’. Someone might ask ‘How certain are you that the baby is exactly 6 kg?’ You might then say ‘Well, the baby might not be exactly 6 kg, but I’m very confident that it weighs somewhere between 5.5 and 6.5 kg’. This statement expresses a confidence interval – a range of values that probably include the real value. Of course, the more certain or confident you want to be of including the true value, the bigger the range of values you might give: you are unlikely to be wrong if you guess that the baby weighs between 4 and 8 kg.

Confidence intervals can be calculated for a large range of statistics such as proportions, ratios or correlation coefficients. In this chapter we will look

at confidence intervals for single samples as an illustration for using confidence intervals.

We have seen previously that if we know the population parameters we can estimate the probability of selecting from that population a sample mean of a given magnitude. Conversely, if we know the sample mean we can estimate the population parameters from which the sample might have come, at a given level of probability. Let us take an example to illustrate this point.

A researcher is interested in the systolic blood pressure (BP) levels of smokers of more than 10 cigarettes per day. She takes a random sample of 100 smokers of 10+ cigarettes in her district and finds the mean BP = 148 mm Hg for the sample, with a standard deviation of $s = 10$.

She wants to generalize to the population of smokers of more than 10 cigarettes per day in their district. The best estimate of μ (the population parameter) is 148, but it is possible that, because of sampling error, 148 is not the exact population parameter. However she can calculate a confidence interval (a range of BPs that will include the true population mean at a given level of probability). A confidence interval is a range of scores which includes the true population parameter at a specified level of probability. The precise probability is decided by the researcher and indicates how certain she can be that the population mean is actually within the calculated range. Common confidence intervals used in statistics are 95% confidence intervals, which offer a probability of $p = 0.95$ for including the true population mean, and 99% confidence intervals, which include the true population mean at a probability of $p = 0.99$.

Calculating the confidence interval requires the use of the following formula:

$$\bar{X} - zS_{\bar{X}} < \mu < \bar{X} + zS_{\bar{X}}$$

where $\bar{x}$ is the sample mean; z is the z -score obtained from the normal curve table such that it cuts off the area of the normal curve corresponding to the

required probability; s_x is the sample standard error, which is equal to the sample standard deviation divided by $\sqrt{n}$ (i.e. $s_x = s / \sqrt{n}$).

Let us turn to the previous example to illustrate the use of this equation. Here $\bar{x} = 148$, and $s_x = 10 / \sqrt{100}$. Assume that we want to calculate a 95% confidence interval. We are looking for a pair of z -scores which have 95% of the standard normal curve between them. In this case, 1.96 is the value for z which cuts off 95% of a normal distribution. That is, we looked up the value of z corresponding to an area (probability) of 0.4750, since the 0.05 has to be divided between the two tails of the distribution, giving 0.025 at either end. Substituting into the equation above we have:

$$148 - (1.96) \left(10 / \sqrt{100} \right) \leq \mu \leq 148 + (1.96) \left(10 / \sqrt{100} \right) = 146.04 \leq \mu \leq 149.96$$

That is, the investigator is 95% confident that the true population mean, the true mean BP of smokers, lies between 146.04 (lower limit) and 149.96 (upper limit). There is only a 5% or 0.05 probability that it lies outside this range. If we chose a 99% confidence interval, then using the formula as above, we have:

$$148 - (2.58) \left(10 / \sqrt{100} \right) \leq \mu \leq 148 + (2.58) \left(10 / \sqrt{100} \right) = 145.42 \leq \mu \leq 150.58$$

Here, 2.58 is the value of z which cuts off 99% of a normal distribution (Fig. 18.6). That is, the investigator is 99% confident that the true population mean lies somewhere between 145.42 and 150.58. Clearly, the 99% interval is wider than the 95% interval; after all, here the probability of including the true mean is greater than for the 95% interval. Conventionally, health sciences publications report 95% confidence intervals.

Confidence Intervals Using the t Distribution

It was previously stated that ‘as n becomes large, the distribution of sample means approaches a normal distribution’ (central limit theorem). The questions left to explain are:

- How large must the sample size, n , be before the sampling distribution of the mean can be considered a normal curve?
- What are the implications for the sampling distribution if n is small?

It has been shown by mathematicians that the sample size, n , for which the sampling distribution of the mean can be considered an approximation of a normal distribution is $n > 30$. That is, if n is 30 or more, we can use the standard normal curve to describe the sampling distribution of the mean. However, when $n < 30$, the sampling distribution of the mean is a rather rough approximation to the normal distribution. Instead of using the normal distribution, we use the t *distribution*, which takes into account the variability of the shape of the sampling distribution due to low n .

The t distribution (Fig. 18.7) is a family of curves representing the sampling distributions of means drawn from a population when sample size, n , is small ($n < 30$). A ‘family of curves’ means that the shape of the t distribution varies with sample size. It has been found that the distribution is determined by the *degrees of freedom* (df) of the statistic.

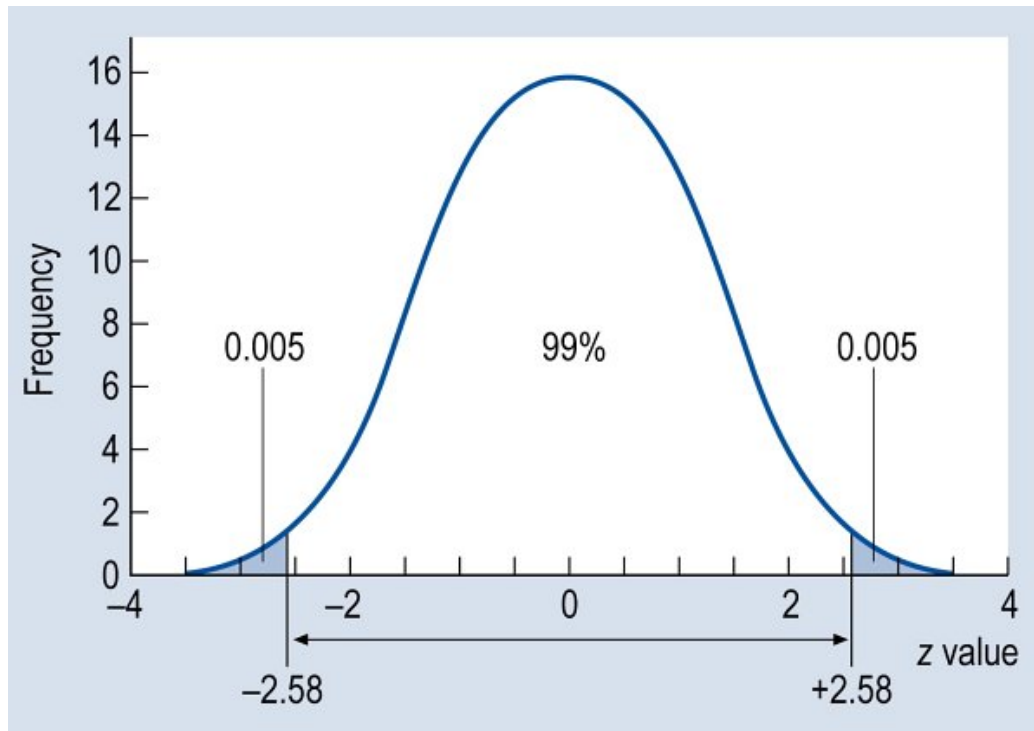


FIG. 18.6 z-scores for 99% confidence interval.

The df for a statistic represents the number of scores which are free to vary when calculating the statistic. Since the statistic we are calculating in this case is the mean, all but one of the scores could vary. That is, if you were inventing scores in a sample with a known mean, you would have a free hand until the very last score. There df is equal to $n - 1$ (the sample size minus one). Each row of figures shown in [Appendix B](#) represents the critical values of t for a given distribution.

1. The t distribution is symmetrical about the mean.
2. The values of t along the x -axis cut off specific areas under the curve, just as for z . These areas are given at the top of the page in [Appendix B](#), under 'Directional' and 'Non-directional' probabilities.
3. The t distribution approaches a normal distribution as n becomes larger. As stated earlier, when $n > 30$, for all practical purposes the t and z distributions coincide.

The t distribution, just as the z distribution, can be used to approximate the probability of drawing sample means of a given magnitude from a population; t can also be used for calculating confidence intervals. Let us reexamine the example relating to the BP of smokers presented earlier. Let us assume that $n = 25$, with the other statistics remaining the same: $\bar{x} = 148$, $s = 10$. The general formula for calculating the confidence intervals for small samples is:

$$\bar{X} - ts_{\bar{x}} < \mu < \bar{X} + ts_{\bar{x}}$$

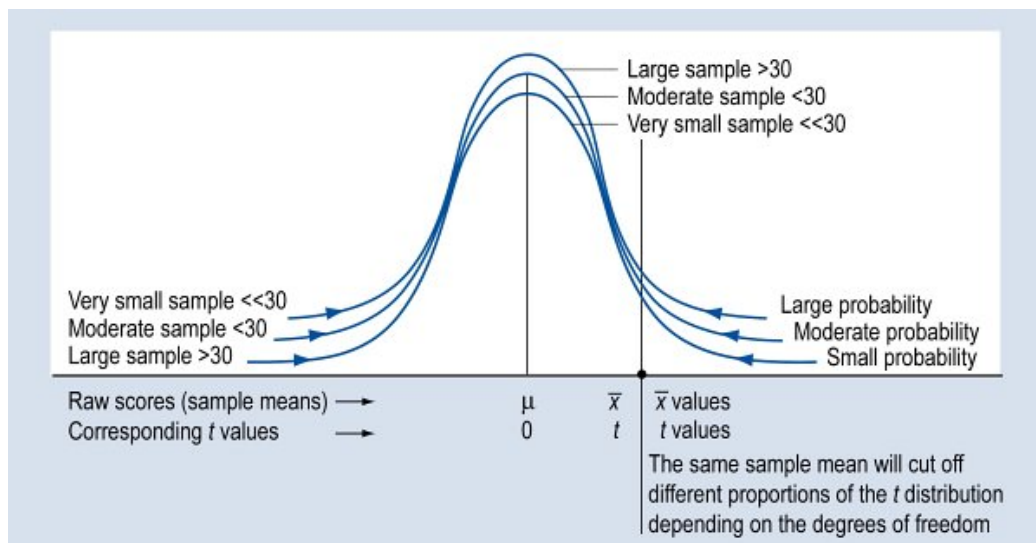


FIG. 18.7 t distributions.

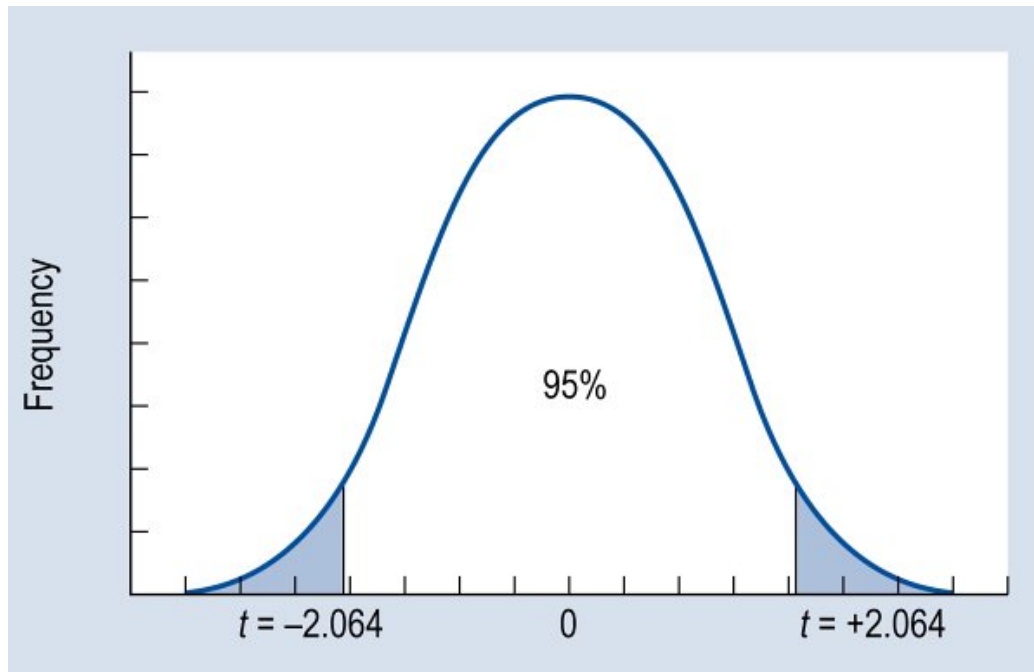


FIG. 18.8 The 95% confidence interval for sample size of 25.

You will note the similarity to the equation on p. 128; here t replaces z . If we want to show the 95% confidence interval, then we use the same logic as for z distributions (Fig. 18.8).

To look up the t values from the tables (Appendix B) consider (1) direction, (2) probabilities and (3) df.

We are looking at a ‘nondirectional’ or ‘two-tail’ probability in the sense that the t values cut off 95% of the area of the t curve between them, leaving 5% distributed at the two tails of the t distribution: $p = 0.05$; $df = 25 - 1 = 24$. Therefore $t = 2.064$ (from Appendix B). Substituting into the equation for calculating confidence intervals:

$$\begin{aligned}
 148 - (2.1) (10/\sqrt{25}) &\leq \mu \leq 148 + (2.1) (10/\sqrt{25}) \\
 148 - 4.2 &\leq \mu \leq 148 + 4.2 \\
 143.8 &\leq \mu \leq 152.2
 \end{aligned}$$

Consider the width of the confidence interval defined as the distance between the upper and lower limits. Note that this is a wider interval than that which was obtained when n was 100. As sample size, n , becomes

smaller, our confidence interval becomes wider, reflecting a greater probability of sampling error.

To calculate the 99% confidence interval (Fig. 18.9) we need to look up $p = 0.01$, nondirectional, $df = 24$ in [Appendix B](#) to obtain the critical value of t , which is 2.797.

$$148 - (2.8) (10/\sqrt{25}) \leq \mu \leq 148 + (2.8) (10/\sqrt{25})$$
$$142.4 \leq \mu \leq 153.6$$

We can see that, when $n = 25$, the 99% confidence interval is wider than for $n = 100$. That is, the bigger our sample size, the narrower (more precise) our estimate of the range of values (which includes the true population parameter) becomes when our sample size is large.

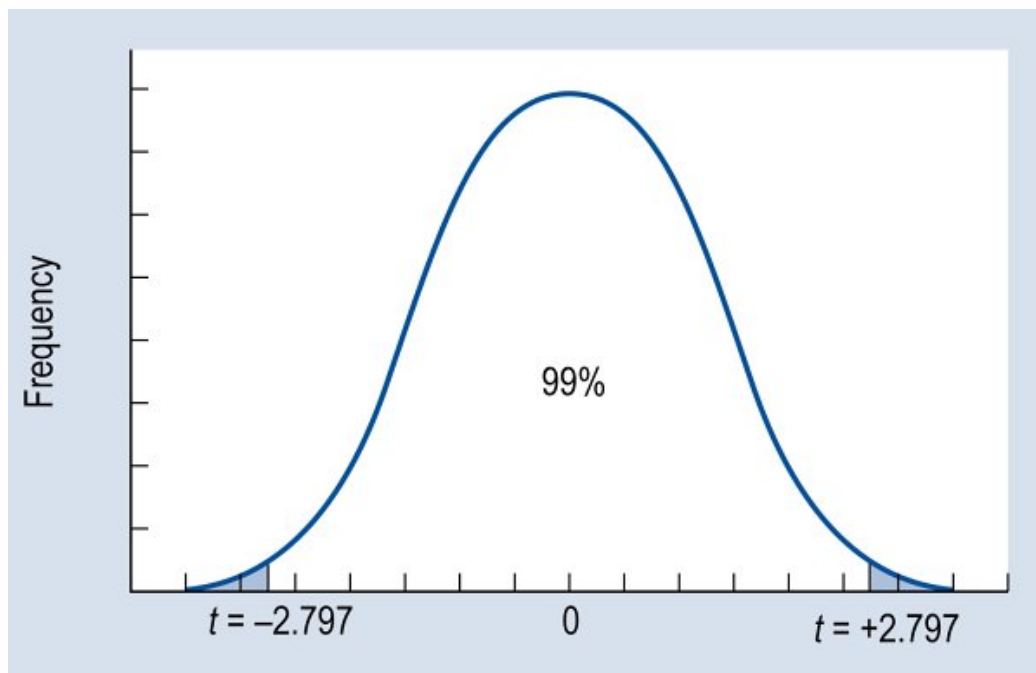


FIG. 18.9 The 99% confidence interval for sample size of 25.

Confidence Intervals in Health Research

Confidence intervals can be calculated for all statistics obtained from sample data. For example, we can work out confidence intervals for proportions, percentages, odds ratios and other measures of effect size. In all these cases, the confidence interval represents an interval that includes at a given probability the true population value or parameter. To calculate the confidence interval we need to know the sampling distribution appropriate for the statistic and use this distribution to calculate the probable upper and lower values of the confidence interval. In general we use a 95% confidence interval which researchers abbreviate as 95% CI.

As an example, consider the study by Mostofsky et al, which investigated the association between alcohol consumption and the relative risk of stroke.

¹ The authors found that the risk for stroke 1 hour after consuming alcohol was 2.3 (95% CI, 1.4–4.0; $p = 0.002$). These statistics indicate that:

1. The risk for stroke in those in the sample who consumed alcohol was 2.3 times the risk for stroke in those who did not have a drink.
2. We can be 95% confident that the population risk of consuming alcohol is between 1.4 and 4. In other words, the true value could be as little as 1.4 or as high as four times the risk for those who drink.
3. The value $p = 0.002$ represents the probability that the risk found in the sample is due to chance. In this case this probability is very small and the results are statistically significant, as we will discuss in [Chapter 19](#).

Confidence intervals are of particular importance in evidence-based practice. The researcher or practitioner is able to see the range of values which contain the true value at a given level of probability, usually 95%. There are numerous websites available which you may want to consult for further information regarding the use and calculation of confidence intervals. This interval estimation is essential for determining the impact of a risk factor on health or the efficacy of a particular intervention. In general, a small sample size combined with highly variable results produces a wide confidence interval. Wide confidence intervals are undesirable in health research and practice because we are unable to be certain where the true population parameter lies. For example, if we say the percentage of patients

improving following a treatment is 50% with a 95% CI of 25% to 75%, then we are very doubtful about the expected outcome. On the other hand, if we could say 50% with a 95% CI of 48% to 52% then we could be sure about the true benefit of the treatment. The narrower the 95% CI, the more accurate the inferences regarding actual population parameters.

Summary

It was argued in this chapter that, even with randomly selected samples, the possibility of sampling error must be taken into account when making inferences from sample statistics to population parameters. It was shown that probability theory could be applied to generating sampling distributions, which express the probability of obtaining a given sample from a population. With discontinuous, nominal data the binomial theorem provides a mathematical distribution for estimating the probability of obtaining possible samples. However, with continuous data, the central limit theorem is applied to generate the sampling distributions of the mean. The standard distribution of the mean enables the calculation of the probability-specified sample mean(s) by random selection. The sampling error of the mean ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$), which expresses statistically the range of the sampling error, depends inversely on the sample size, such that the larger the n , the smaller the $s_{\bar{x}}$ or $\sigma_{\bar{x}}$.

One of the applications of sampling distributions is for calculating confidence intervals for continuous data. Confidence intervals represent a range of scores which specify, from sample data, the probability of capturing the true population parameters. Confidence intervals are essential statistics for the conduct of health research and evidence-based health care.

Health researchers usually report 95% confidence intervals. When sample sizes are small ($n < 30$), the t distribution is appropriate for calculating the sampling distribution of the mean. With large sample sizes, the two distributions merge together. As the next chapter will demonstrate, sampling distributions are essential for testing hypotheses, a procedure which uses inferential statistics to calculate the level of probability at which sample statistics support the predictions of hypotheses.

Reference

1. Mostofsky E, Burger M, Schlaug G. Alcohol and acute ischemic stroke onset: the stroke onset study. *Stroke* . 2010;41(9):1845–1849.

Hypothesis Testing

Introduction,
The Logic of Hypothesis Testing,
Steps in Hypothesis Testing,
Illustrations of Hypothesis Testing,
The Relationship Between Descriptive and Inferential
Statistics,
Selection of the Appropriate Inferential Test,
The χ^2 Test,
 χ^2 and Contingency Tables,
Statistical Packages,
Summary,

Introduction

Hypotheses are statements about the association between variables as pertaining to a specific person or population – for example, ‘penicillin is an effective treatment for pneumonia’ or ‘obesity is a risk factor for heart disease’. Hypotheses addressing the state of populations are tested using sample data. *Inferences* are conclusions based on data using samples and

are therefore always open to the possibility of error. In this chapter, we will examine the use of inferential statistics for establishing the probable truth of hypotheses, as tested through sample data. Inferential statistics are based on applied probability theory and entail the use of statistical tests.

There are numerous statistical tests available that are used in a similar fashion to analyse clinical data. That is, all statistical tests involve setting up the relevant hypotheses, H_0 and H_A , and then, on the basis of the appropriate inferential statistics, computing the probability of the sample statistics obtained occurring by chance alone. We are not going to attempt to examine all statistical tests in this introductory book. These are described in various statistics textbooks or in data analysis manuals. Rather, in this chapter, we will examine the criteria used for selecting tests appropriate for the analysis of the data obtained in specific investigations. To illustrate the use of statistical tests we will examine the use of the chi-square test (χ^2). This is a statistical test commonly employed to analyse categorical data. Finally, we will briefly examine the uses of the SPSS Version 26 for data analysis in general.

The aims of this chapter are to:

1. Discuss the criteria by which a statistical test is selected for analysing the data for a specific study.
2. Demonstrate the use of the χ^2 test for analysing nominal scale data.
3. Explain how statistical packages are used for quantitative data analysis.

The Logic of Hypothesis Testing

Hypothesis testing is the process of deciding to use statistics whether the findings of an investigation reflect chance or real effects at a given level of probability or certainty. If the results do not seem to represent chance effects, then we say that the results are statistically significant. That is, when we say that our results are statistically significant, we mean that the patterns or differences seen in the sample data are likely to be generalizable to the wider population from our study sample.

The mathematical procedures for hypothesis testing are based on the application of probability theory and sampling, as discussed previously. Because of the probabilistic nature of the process, decision errors in hypothesis testing cannot be entirely eliminated. However, the procedures outlined in this chapter enable us to specify the probability level at which we can claim that the data obtained in an investigation support experimental hypotheses. This procedure is fundamental for determining the statistical significance of the data and being relevant to the logic of clinical decision making.

Steps in Hypothesis Testing

The following steps are conventionally followed in hypothesis testing:

1. State the alternative hypothesis (H_A), which is based on the research hypothesis. The H_A asserts that the results are ‘real’ or ‘significant’ (i.e. that the independent variable influenced the dependent variable, or that there is a real difference among groups). The important point here is that H_A is a statement concerning the population. A real or significant effect means that the results in the sample data can be generalized to the population.
2. State the null hypothesis (H_0), which is the logical opposite of the H_A . The H_0 claims that any differences in the data were just due to chance: that the independent variable had no effect on the dependent variable, or that any difference among groups is due to random effects. In other words, if the H_0 is retained, differences or patterns seen in the sample data should not be generalized to the population.
3. Set the decision level, α (alpha). There are two mutually exclusive hypotheses (H_A and H_0) competing to explain the results of an investigation. Hypothesis testing, or statistical decision making, involves establishing the probability of H_0 being true. If this probability is very small, we are in a position to reject the H_0 . You might ask ‘How small should the probability (α) be for rejecting

- H_0 ?' By convention, we use the probability of $\alpha = 0.05$. If the H_0 being true is less than 0.05, we can reject H_0 . We can choose an α of 0.05, but not more. That is, by convention among researchers, results are not characterized as significant if $p > 0.05$.
4. Calculate the probability of H_0 being true. That is, we assume that H_0 is true and calculate the probability of the outcome of the investigation being due to chance alone (i.e. due to random effects). We must use an appropriate sampling distribution for this calculation.
 5. Make a decision concerning H_0 . The decision rule used is as follows: if the probability of H_0 being true is less than α , then we reject H_0 at the level of significance set by α . However, if the probability of H_0 is greater than α , then we must retain H_0 . In other words, if:
 - a. $p(H_0 \text{ is true}) \leq \alpha$, reject H_0
 - b. $p(H_0 \text{ is true}) > \alpha$, retain H_0

It follows that if we reject H_0 , we are in a position to accept H_A , its logical alternative. If $p \leq 0.05$ then we reject H_0 , and decide that H_A is probably true.

Illustrations of Hypothesis Testing

One of the simplest forms of gambling is betting on the fall of a coin. Let us play a little game. We, the authors, will toss a coin. If it comes out heads (H) you will give us ≤ 1 ; if it comes out tails (T) we will give you ≤ 1 . To make things interesting, let us have 10 tosses. The results are:

Toss	1	2	3	4	5	6	7	8	9	10
Outcome	H	H	H	H	H	H	H	H	H	H

Oh dear, you seem to have lost. Never mind, we were just lucky, so send along your cheque for ≤ 10 . Are you a little hesitant? Are you saying that we

‘fixed’ the game? There is a systematic procedure for demonstrating the probable truth of your allegations:

1. We can state two competing hypotheses concerning the outcome of the game:
 - a. The authors fixed the game; that is, the outcome did not reflect the fair throwing of a coin. Let us call this statement the ‘alternative hypothesis’, H_A . In effect, the H_A claims that the sample of 10 heads came from a population other than P (probability of heads) = Q (probability of tails) = 0.5
 - b. The authors did not fix the game; that is, the outcome is due to the tossing of a fair coin. Let us call this statement the ‘null hypothesis’, or H_0 . H_0 suggests that the sample of 10 heads was a random sample from a population where $P = Q = 0.5$.
2. It can be shown that the probability of tossing 10 consecutive heads with a fair coin is actually $p = 0.001$, as discussed previously (see [Ch. 18](#)). That is, the probability of obtaining such a sample from a population where $P = Q = 0.5$ is extremely low.
3. Now we can decide between H_0 and H_A . It was shown that the probability of H_0 being true was $p = 0.001$ (1 in a 1000). Therefore in the balance of probabilities, we can reject it as being true and accept H_A , which is the logical alternative. In other words, it is likely that the game was fixed and no ≤ 10 cheque needed to be posted.

The probability of calculating the truth of H_0 depended on the number of tosses (n = the sample size). For instance, the probability of obtaining heads every time with five coin tosses is shown in [Table 19.4](#). As the sample size (n) becomes larger, the probability for which it is possible to reject H_0 becomes smaller. With only a few tosses we really cannot be sure if the game is fixed or not: without sufficient information it becomes hard to reject H_0 at a reasonable level of probability.

A question emerges: ‘What is a reasonable level of probability for rejecting H_0 ?’ As we shall see, there are conventions for specifying these probabilities. One way to proceed, however, is to set the appropriate probability for rejecting H_0 on the basis of the implications of erroneous decisions.

Obviously, any decision made on a probabilistic basis might be in error. Two types of decision errors are identified in statistics as *type I* and *type II errors*. A type I error involves mistakenly rejecting H_0 , whereas a type II error involves mistakenly retaining the H_0 . Researchers can make mistakes about the truth or falsity of hypotheses using sample research data. The statistical method does not provide a guarantee against making a mistake, but it is the most rigorous way of making these decisions.

In the coin toss example, a type I error would involve deciding that the outcome was not due to chance when in fact it was. The practical outcome of this would be to falsely accuse the authors of fixing the game. A type II error would represent the decision that the outcome was due to chance, when in fact it was due to a ‘fix’. The practical outcome of this would be to send your hard-earned ≤ 10 to a couple of crooks. Clearly, in a situation like this, a type II error would be more odious than a type I error, and you would set a fairly high probability for rejecting H_0 . However, if you were gambling with a villain who had a loaded revolver handy, you would tend to set a very low probability for rejecting H_0 . We will examine these ideas more formally in subsequent parts of this chapter.

Let us look at another example. A rehabilitation therapist has devised an exercise program that is expected to reduce the time it takes for people to leave hospital after orthopaedic surgery. Previous records show that the recovery time for patients had been $\mu = 30$ days, with $\sigma = 8$ days. A sample of 64 patients were treated with the exercise program, and their mean recovery time was found to be $\bar{x} = 24$ days. Do these results show that patients who had the treatment recovered significantly faster than previous patients? We can apply the steps for hypothesis testing to make our decision.

1. State H_A : 'The exercise program reduces the time taken for patients to recover from orthopaedic surgery'. That is, the researcher claims that the independent variable (the treatment) has a 'real' or 'generalizable' effect on the dependent variable (time to recover).
2. State H_0 : 'The exercise program does not reduce the time taken for patients to recover from orthopaedic surgery'. That is, the statement claims that the independent variable has no effect on the dependent variable. The statement implies that the treated sample with $\bar{x} = 24$, and $n = 64$ is in fact a random sample from the population $\mu = 30$, $\sigma = 8$. Any difference between $\bar{x}$ and μ can be attributed to sampling error.
3. The decision level, α , is set before the results are analysed. The probability of α depends on how certain the investigator wants to be that the results show real differences. If he sets $\alpha = 0.01$, then the probability of falsely rejecting a true H_0 is ≤ 0.01 (1/100). If he sets $\alpha = 0.05$, then the probability of falsely rejecting a true H_0 is ≤ 0.05 or (1/20). That is, the smaller the α , the more confident the researcher is that the results support the alternative hypothesis. We also call α the level of significance. The smaller the α , the more significant the findings for a study, if we can reject H_0 . In this case say that the researcher sets $\alpha = 0.01$. (Note: by convention, α should not be greater than 0.05.)

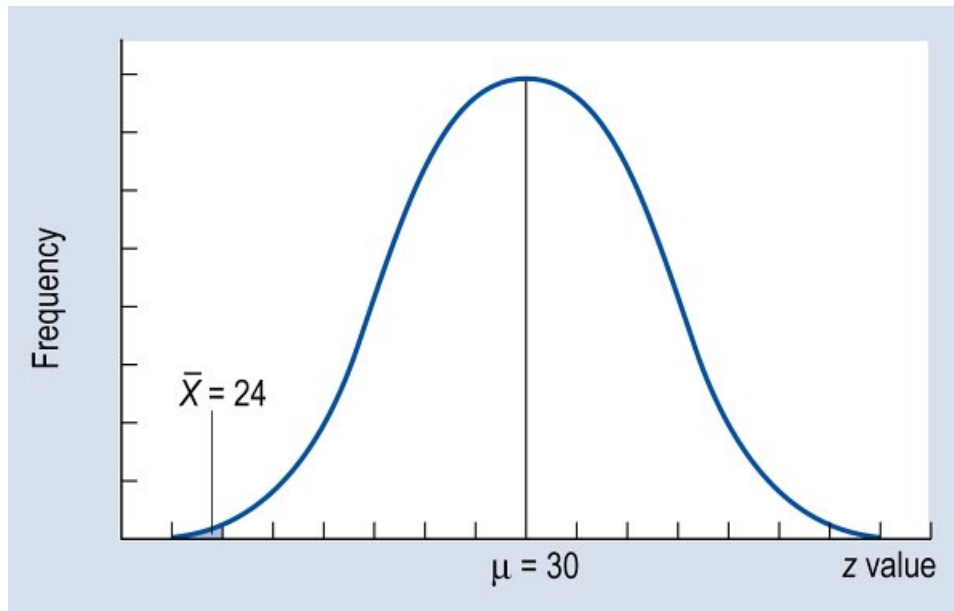


FIG. 19.1 Sampling distribution of means. Sample size = 64, population mean = 30, standard deviation = 8.

4. Calculate the probability of H_0 being true. As stated previously, H_0 implies that the sample with $\bar{x} = 24$ is a random sample from the population with $\mu = 30$, $\sigma = 8$. How probable is it that this statement is true? To calculate this probability, we must generate an appropriate sampling distribution. The sampling distribution of the mean will enable us to calculate the probability of obtaining a sample mean of $\bar{x} = 24$ or more extreme from a population with known parameters. As shown in Fig. 19.1, we can calculate the probability of drawing a sample mean of $\bar{x} = 24$ or less. Using the table of normal curves (Appendix A), we find that the probability of randomly selecting a sample mean of $\bar{x} = 24$ (or less) is extremely small. In terms of our table, which only shows the exact probability of up to $z = 4.00$, we can see that the present probability is <0.00003 . Therefore the probability that H_0 is true is <0.00003 .
5. Make a decision. We have set $\alpha = 0.01$. The calculated probability was <0.0001 . Clearly, the calculated probability is far less than α , indicating that the difference is unlikely to be due to chance. Therefore the investigator can reject the statement that H_0 is true

and accept H_A , that patients in general treated with the exercise program recover earlier than the population of untreated patients.

The Relationship Between Descriptive and Inferential Statistics

As we have seen in the previous chapters, statistics may be classified as descriptive or inferential. Descriptive statistics describe the characteristics of data and are concerned with issues such as ‘What is the average length of hospitalization of a group of patients?’ Inferential statistics are used to address issues such as whether the differences in average lengths of hospitalization of patients in two groups are significantly different statistically. Thus descriptive statistics describe aspects of the data such as the frequencies of scores, and the average or the range of values for samples, whereas inferential statistics enable researchers to decide (infer) whether differences between groups or relationships between variables represent persistent and reproducible trends in the populations.

We have seen that the selection of appropriate descriptive statistics depends on the type of data being described. For example, in a variable such as incomes of patients, the best statistics to represent the typical income would be the mean and/or the median. If you had a millionaire in the group of patients, the mean statistic might give a distorted impression of the central tendency. In this situation the median statistic would be the most appropriate one to use. The mode is most commonly used when the data being described are categorical data. For example, if in a questionnaire respondents were asked to indicate their sex and 65% said they were male and 35% said they were female, then ‘male’ is the modal response. It is quite unusual to use the mode only with data that are not nominal. As a rule, the scale of measurement used to obtain the data and its distribution determine which descriptive statistics are selected.

In the same way, the appropriate inferential statistics are determined by the characteristics of the data being analysed. For example, where the mean is the appropriate descriptive statistic, the inferential statistics will determine whether the differences between the means are statistically significant. In the case of ordinal data, the appropriate inferential statistics

will make it possible to decide if either the medians or the rank orders are significantly different. With nominal data, the appropriate inferential statistic will decide if proportions of cases falling into specific categories are significantly different.

Thus when the data have been adequately described, the appropriate inferential statistic will follow logically. However, when selecting an appropriate statistical test, the design of the investigation must also be taken into account.

Selection of the Appropriate Inferential Test

Before addressing the issue of the selection of the appropriate inferential statistical test, it is useful to reiterate the reason why a statistical test should be employed in research.

In some studies, inferential statistical tests are not required. For example, if a health care needs assessment survey is conducted in a particular community, using a full population, the investigator might not be overly concerned with generalizing the results to other communities, or demonstrating that certain relationships between variables are reliable. It may be enough to be able to say, for example, that ‘35% of the respondents indicated that they were dissatisfied with the existing level of medical services’. In this instance, descriptive statistics are all that the investigator requires to do the job because the complete population was studied. If, however, the investigator wishes to argue that certain differences between groups or certain correlations between variables for a sample are generalizable to the population, then inferential statistical tests are necessary.

The inferential statistic provides the investigator with a means of determining how reproducible the obtained results are, by enabling access to a probability. The probability associated with the value of an inferential statistic informs the investigator of the likelihood that the results obtained were due to chance factors, or if they are significant at a given level of probability.

As noted earlier, we are not going to examine all of the numerous statistical tests available for decision making. Rather, the aim of this chapter is to examine the criteria used for selecting tests appropriate for the analysis

of data obtained in investigations. To illustrate the use of statistical tests we will look at the χ^2 test, commonly employed for analysing nominal data. We examine the interpretation of findings that do not reach statistical significance and the relationship between statistical and clinical significance. In [Chapter 21](#) we will consider some of the determinants of making decisions concerning the actual adoption and use of treatments and diagnostic tests in clinical settings.

There are a variety of statistical tests, some of which are shown in [Table 19.1](#). The selection of the appropriate statistical test is determined by four considerations:

1. The scale of measurement used to obtain the data (nominal, ordinal, interval, or ratio).
2. The number of groups used in an investigation (one or more).
3. Whether the measurements were obtained from independent participants or from related samples, such as those involving repeated measurements of the same participants.

Table 19.1

Scale	Two groups		Three or more groups	
	Independent	Dependent	Independent	Dependent
Nominal	χ^2 test	McNemar's test	χ^2 test	Cochran's Q test
Ordinal	Mann-Whitney <i>U</i> test	Sign test	Kruskal-Wallis <i>H</i> test	Friedman two-way ANOVA
Interval or ratio	<i>t</i> test (independent groups)	<i>t</i> test (dependent groups)	ANOVA (<i>F</i>) (independent groups)	ANOVA (<i>F</i>) (dependent groups)

ANOVA, analysis of variance.

4. The assumptions involved in using a statistical test, such as the distribution of the scores or the minimum required sample size.

[Table 19.1](#) shows a number of statistical tests to illustrate how statistical tests are selected for analysing data. Several points are worth noting:

1. It can be seen that appropriate tests are selected on the basis of the four criteria just outlined. When we have determined these four criteria for a given investigation, the 'cell' containing the appropriate test can be readily selected. Although not shown in [Table 19.1](#), there are instances when there are a number of

statistical tests available in a particular cell. Here we need additional criteria for deciding between two tests within a cell. For instance, we saw in the previous section that, if $n < 30$, we use the t test rather than the z test.

2. The tests appropriate for analysing ordinal and nominal data are nonparametric, or distribution free. The tests for analysing interval or ratio data are called parametric tests. The parametric tests (e.g. z , t or F) require that certain assumptions (such as normality and equal variance) be valid for the populations from which the samples were drawn. The nonparametric tests (e.g. χ^2 , Mann–Whitney U) require few, if any, assumptions about the underlying population distributions.
3. Even before the data are collected, an investigator should have a good idea of which statistical test is appropriate for analysing the data. Sometimes, however, the distribution of the data is such that the test that was initially selected is found to be inappropriate.

Let us look at some examples to illustrate how statistical tests are selected.

An investigator wishes to evaluate the effectiveness of a new treatment in contrast to a conventionally used treatment. Assume that the outcome (dependent variable) is measured on a 5-point ordinal scale. Each participant is assigned to one of the two treatment groups. Which test would the investigator use to analyse the significance of sample data when:

1. The measurement was ordinal?
2. There were two groups (new treatment, conventional treatment)?
3. Participants were independently assigned to a specific group?

By inspection of [Table 19.1](#) the investigator would select the Mann–Whitney U test to analyse the significance of the data.

If we change this example by stating that the dependent variable was measured on an interval scale, the appropriate test would now be a t test (for independent groups). Let us say that three groups were used (by the inclusion of a placebo group) by the investigator. Now, if the outcome measurement remained ordinal, the appropriate test for analysing the data is the Kruskal–Wallis H test. If, however, the outcome measures were interval,

it follows from [Table 19.1](#) that the appropriate test for analysing the results would be the analysis of variance.

Finally, say that in the original example each of the participants was treated with both the new and old treatments. Now, the data would have been obtained from the repeated measurement of the same participants, and the appropriate statistical test would be the Sign test (ordinal, two groups, dependent).

[Table 19.1](#) does not include all the available statistical tests and their uses. In fact, mathematical statisticians can generate inferential tests appropriate for a whole variety of designs. The basic idea is to use probability theory to generate appropriate sampling distributions in terms of whether the probability of H_0 being true can be calculated, and the statistical significance of the findings evaluated. ¹

Rather than examining all the tests and their underlying assumptions, we will look at the use of the χ^2 test in some detail. In addition to being a very useful test for analysing nominal data, it (along with the z and t) illustrates how statistical tests are carried out to test hypotheses.

The χ^2 Test

As shown in [Table 19.1](#), χ^2 is appropriate for statistical analysis when:

- Variables are measured on a nominal scale.
- Measurements are of independent participants.

Table 19.2

Treatments Offered

Psychotherapy	Drugs	Electroconvulsive Therapy
$n = 45$	$n = 40$	$n = 65$

The χ^2 test is appropriate for deciding if proportions of cases falling into categories are different at a given level of significance.

The statistic, χ^2 , is given by the formula:

$$\chi^2 = \frac{\sum (f_o - f_e)^2}{f_e}$$

where f_o = observed frequency for a given category and f_e = expected frequency for a given category, assuming H_0 was true.

The sampling distribution for χ^2 is a family of curves, which, like t , vary with degrees of freedom. The use of this inferential statistic is best illustrated by an example.

Suppose that an investigator is interested in finding out whether there is a difference in the relative frequency of different kinds of treatments currently offered to extremely depressed patients. A random sample of 150 patients is selected from a population of patients in Australia, and the type of treatment offered to them is determined from their medical records, as shown in [Table 19.2](#).

The entries in each cell represent the frequency with which patients were given the various treatments. Thus 45 patients were offered psychotherapy, 40 patients offered drugs and 65 patients offered electroconvulsive therapy. The χ^2 is the appropriate test for analysing these data. Let us follow the steps involved in hypothesis testing, as outlined previously.

1. H_A : there is a difference in the population proportions for the three treatments. H_A is nondirectional when we use χ^2 .
2. H_0 : there is no difference in the population proportions of the three treatments. The frequencies shown in each cell in the table occurred through random sampling from a population where there is an equal frequency of the three treatments.
3. Decision level, α : say the investigator sets a significance level of 0.05 for rejecting H_A ($\alpha = 0.05$).
4. Calculation of the statistic: χ^2_{obt} is the value of χ^2 calculated from the data obtained. To calculate χ^2_{obt} , we must determine f_e for each cell (f_o is, of course, determined by the data). If the null hypothesis (H_0) is true, then our expectation is that the frequencies in each cell

should be the same. In this case $n = 150$, so that f_e should be $150/3 = 50$, given that there are three cells. Let us show this in tabular form (Table 19.3). We can now calculate χ^2_{obt} by calculating $(f_o - f_e)^2/f_e$ for each cell, and then summing the values:

Table 19.3

Treatments (f_e is shown in parentheses)

Psychotherapy	Drugs	Electroconvulsive Therapy
45	40	65
(50)	(50)	(50)

$$\begin{aligned}
 \chi^2_{\text{obt}} &= \sum \frac{(f_o - f_e)^2}{f_e} \\
 &= \frac{(45-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(65-50)^2}{50} \\
 &= 0.5 + 2.0 + 4.5 \\
 &= 7.0
 \end{aligned}$$

The greater the discrepancy between f_e and f_o , the greater the calculated value of the chi-square statistic (χ^2_{obt}). The direction of the difference is of no significance as the difference between f_e and f_o is squared.

5. Making decisions concerning H_0 : the decision rule for χ^2 is similar to that of the z and t tests, as shown in the previous chapter:

$$\chi^2_{\text{obt}} \geq \chi^2_{\text{crit}}, \text{ reject } H_0$$

$$\chi^2_{\text{obt}} < \chi^2_{\text{crit}}, \text{ retain } H_0$$

Here, χ^2_{crit} is the critical value of the statistic χ^2 , which cuts off a proportion of the sampling distribution equal to α . The value of χ^2_{crit} is obtained from the tables in [Appendix C](#). To look up this statistic, we need to know:

- a. α , which was set at 0.05 for this example
- b. the degrees of freedom, df

Note that with χ^2 the df with one variable is $k - 1$, where k stands for the number of categories or groups. In this instance we have $k = 3$ (three treatments) so that $df = 3 - 1 = 2$. Now we can look up the tables in [Appendix C](#). In this case $\alpha = 0.05$ and $df = 2$, therefore $\chi^2_{\text{crit}} = 5.99$.

Here, because $\chi^2_{\text{obt}} > \chi^2_{\text{crit}}$, we can reject H_0 at a 0.05 level of significance. The investigator is in a position to accept H_A (that the three treatments are offered at different frequencies to depressed patients). Clearly, electroconvulsive therapy is given most frequently for the condition (in this hypothetical example).

χ^2 and Contingency Tables

In the previous example of χ^2 , we had clear expectations of the expected frequencies (f_e) and were dealing with only one variable. The χ^2 test is also relevant for analysing nominal data where f_e is not known and where we are interested in the effects of more than one variable. Thus χ^2 is a statistical test appropriate for deciding whether two variables are significantly related.

Table 19.4

	Drugs	Surgery	Row Marginal
Dead	11	8	19
Alive	9	12	21
Column marginal	20	20	$n = 40$

For example, an investigator compares the effectiveness of drug therapy with coronary artery surgery in men 55 to 60 years old suffering from coronary heart disease. A sample of 40 patients consenting to the investigation is selected from this population and randomly divided into the two treatment groups (drugs only or coronary artery surgery). The treatment outcome is measured in terms of survival over 5 years. The outcome of this hypothetical study is shown in [Table 19.4](#).

[Table 19.4](#) is called a *contingency table*. A contingency table is a two-way table showing the relationship between two or more variables. Note that the levels of the variables have been classified into mutually exclusive categories ('drugs or surgery' for the independent variable, and 'dead or alive' for the dependent variable, in this instance). The cells in the contingency table show the frequency of cases falling into each joint category (e.g. 11 people in the 'drugs only' category had died during the 5 years). The row and column marginal scores are the sums of the frequencies. The row and column marginals necessarily add up to n , the sample size ($n = 40$ for this example).

[Table 19.4](#) is called a 2×2 contingency table. Depending on the number of categories (or levels) in each of the two variables, we might have 3×2 tables, 3×3 tables, and so on. Let us now turn to analysing the data.

1. H_A : there is a difference in the proportion of patients surviving for 5 years after the two types of treatment.
2. H_0 : there is no difference in the frequency of survival rates; any difference between observed and expected frequencies in the sample is due to chance.
3. Decision level: $\alpha = 0.05$
4. Calculation of the statistic χ^2_{obt} :

$$\chi^2_{\text{obt}} = \sum \frac{(f_o - f_e)^2}{f_e}$$

Table 19.5

General Format for 2 × 2 Contingency Table

A	B	<i>j</i>
C	D	<i>k</i>
<i>l</i>	<i>m</i>	<i>n</i>

Table 19.6

	f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
A	11	9.5	2.25	0.236
B	8	9.5	2.25	0.236
C	9	10.5	2.25	0.214
D	12	10.5	2.25	0.214
<i>n</i>	40			$\chi^2 = 0.90$

To make our explanation of the calculation easier, let us label the cells and the marginal values, as shown in [Table 19.5](#). We calculate χ^2_{obt} by calculating f_e for each of the cells and then substituting this value into the equation for χ^2_{obt} . To calculate the expected frequencies, f_e , for each of the cells, we use the formula:

$$\frac{\text{Row total} \times \text{column total}}{n}$$

Substituting into the formula for each of the cells:

$$A: f_e = \frac{j \times l}{n} \quad B: f_e = \frac{j \times m}{n}$$

$$C: f_e = \frac{k \times l}{n} \quad D: f_e = \frac{k \times m}{n}$$

Now f_o are the observed frequencies as in the data, summarized in [Table 19.5](#). Substituting the values for f_e and f_o for each cell is shown in [Table 19.6](#).

5. Making decisions concerning H_0 : the degrees of freedom for a contingency table are calculated by:

$$df = (r - 1)(c - 1)$$

where r = the number of rows and c = the number of columns. In this instance given a 2×2 contingency table:

$$df = (2 - 1)(2 - 1) = 1$$

Now we can look up χ^2_{crit} for $\alpha = 0.05$ and $df = 1$. From [Appendix C](#), $\chi^2_{crit} = 3.84$. Therefore because $\chi^2_{obt} < \chi^2_{crit}$, we must retain H_0 : there is no statistically significant difference in the frequencies of survival over 5 years after the two kinds of treatments. Our data

show that either there is no difference in the outcomes of the two treatments or we made a type II error.

The χ^2 test can be used to analyse the statistical significance of nominal data arising from experimental or nonexperimental investigations. This nonparametric test can be used provided that two simple assumptions are met:

1. Each participant has provided only one entry into the χ^2 table; that is, each of the entries is independent.
2. The expected frequency (f_e) in each cell is at least 5. Therefore if the sample size is too small, χ^2 may not be used.

If either of these assumptions is violated, the use of χ^2 is inappropriate for statistical decision making. Assumption 2 is particularly important when $df = 1$ for a contingency table.

Statistical Packages

We have considered several examples of establishing the statistical significance of the results. These calculations were presented only for teaching purposes. Some older applied statistics textbooks are crammed with complicated formulae for calculating dozens of different inferential statistics. Researchers now use statistical packages which have made statistical analysis simpler and more accessible to all researchers. The following steps are followed when using statistical packages:

1. *Select a statistical package.* There are many packages on the market, including Statistical Package for Social Sciences ('SPSS'), 'STATISTICA', 'Statsview' or various spreadsheets with useful statistical functions. Each program has its strengths and weaknesses, and some researchers have formed strong attachment to specific programs. If you are a beginner, you should be guided by your thesis supervisor or your workplace mentor about the availability of packages.

2. *Training.* It is useful to learn how to use a package before you begin data analysis. Depending on your aptitude and experience, training sessions take 1 or 2 days. With more complex scientific packages such as ‘STATISTICA’ it might require long-term usage before one feels like an expert user.
3. *Encode the raw data.* We begin by encoding the data. You have to be clear about issues such as which are the independent or dependent variables, or the scaling of the data (continuous/discontinuous). That is, designs and measurement procedures used in your research project influence the encoding process.
4. *Identifying the statistical analysis required.* Some degree of statistical knowledge is required beyond that covered in this book to enable you to confidently select and interpret the appropriate statistical analyses. Keep in mind that our book is introductory; it is expected that you will complete a more advanced statistics subject. For more complex analyses it might be useful to seek expert help.

Table 19.7

Treatment	N	Mean	Standard Deviation	Standard Error Mean
Control (C)	11	8.00	2.864	0.863
Exercise (E)	11	9.00	3.821	1.152

5. *Printout.* You will select the appropriate statistical analysis from the ‘menu’ and print out the results of the analysis.
6. *Interpretation.* Finally you will need to interpret the ‘printout’ in relation to your research questions and/or hypotheses.

Let us look at a simple example. Say we are interested in the benefits of exercise for improving mobility in nursing home residents. A sample of 24 ($n = 24$) residents with reduced mobility are selected and have given informed consent to participate. The participants are randomly assigned to either of two groups: ‘E’, which involves them undertaking an exercise program suitable for improving mobility in elderly patients, and ‘C’, which involves undertaking alternative activities of equal duration. The outcome or dependent variable is measured as the distance in metres safely walked by the

residents unassisted at the completion of the exercise and control programs. The research hypothesis here is: exercise improves mobility in nursing home residents. We will look at a set of created data to illustrate hypothesis testing using a statistical package (SPSS).

Let us follow the steps for analysing the data:

1. Assume that the SPSS package is readily available in your workplace and you are informed by your computing expert or supervisor.
2. You have had some training. Even if no formal training is available, self-help books such as *SPSS: analysis without anguish. Version 11.0 for Windows* are very useful. ²
3. After accessing the spreadsheet for the program, we encode the data in the two columns representing 'E' and 'C'. We inform the program that the data are 'continuous' (ratio-scale data).
4. We refer to [Table 19.1](#) to identify the required statistical analysis. Here we have:
 - a. two groups.
 - b. independent groups.
 - c. ratio-scale data.

Therefore we select the independent t test for analysing the data from the 'menu' of inferential tests available.

5. Printouts: Tables [19.7](#) and [19.8](#) for SPSS are based on printouts presented in Coates and Steed.^{2(p73)} We changed the original example which contains additional information: the interested reader might wish to examine it further in Coates and Steed. ²

Table 19.8

Treatment	t	df	Sig. (2-Tailed)	Mean Difference	Standard Error Difference	95% confidence interval	
						Lower	Upper
Control (C)	-0.695	20	0.495	-1.00	1.440	-4.003	2.003
Exercise (E)	-0.695	18.539	0.496	-1.00	1.440	-4.018	2.013

[Table 19.7](#) shows key descriptive statistics:

- n refers to the sample size for each group.
- mean ($\bar{x}$) for each group.
- standard deviation (s) for each group.
- standard error mean (s_x) for each group. $s_x = s / \sqrt{n}$. Here, for the control group, 'C' is the standard error of the mean and is equal to $2.864 / \sqrt{11} = 0.863$, as shown in [Table 19.7](#).

The group mean for the exercise group does in fact indicate a better overall performance. However, we are justified in concluding that exercise in general produces improved mobility in nursing home residents. To answer this question we must look at the results of the t test, as shown in [Table 19.8](#).

The following information was presented:

- The t obtained was -0.695 , the minus sign simply reflecting that the mean for the exercise group was less than the mean for the control group. A t value of 0.695 is quite 'small' in relation to the critical values for t shown in [Appendix B](#).
- 'df' refers to degrees of freedom. For the control group's t test, $df = n_1 + n_2 - 2$. Here, $df = 11 + 11 - 2 = 20$, as shown in [Table 19.8](#).
- 'Sig. (2-tailed)' refers to the probability of a difference between the means being obtained by chance, i.e. $p(H_0 \text{ is true})$. The decision rule is that we reject H_0 if $p < 0.05$. Here the calculated probability is 0.495 and therefore we retain H_0 . We conclude that the results were not statistically significant.
- The 'mean difference' refers to the difference between the two sample means; i.e. $8 - 9 = -1$, as shown in [Table 19.8](#).
- 'Standard error difference' refers to the standard error of the distribution of sample means. We will not discuss this statistic in detail. The standard error enables us to calculate a 95% confidence interval; in this case for $\bar{x}_C - \bar{x}_E$.
- The lower and upper limits show that the true (i.e. population) difference ($\mu_C - \mu_E$) probably ($p = 0.95$) falls between -4.003 and 2.003 . This is a wide range for a confidence interval and therefore indicates that residents who exercise might be able to walk either

four extra metres or two metres less than those who don't exercise. Of course, any other difference between the two limits is possible. Clearly, such results are far too variable to attribute any clinical benefits to the exercise program.

Summary

In this chapter, we examined the use of statistical tests for analysing and interpreting the results of quantitative research. There are a variety of statistical tests available for analysing the significance of the obtained data. The statistical test appropriate for analysing a given set of data is selected on the basis of:

1. The scaling of the data.
2. The dependence/independence of the measurements.
3. The number of groups being studied.
4. Specific requirements for using a statistical test.

Generally, parametric and nonparametric statistical tests were distinguished on the grounds of the scaling of the data and the assumptions underlying the sampling distributions.

None of the individual statistical tests were discussed in detail, except the χ^2 test, which was presented as an example. The χ^2 test illustrates the principle that theoretical sampling distributions can be generated, and the probability of obtaining specific outcomes can be calculated. If the obtained value of the inferential statistic is greater than or equal to the critical value, the null hypothesis can be rejected at the level of significance specified by the type I error rate (α). This is the case regardless of which particular statistical test is being used. The retention of H_0 might reflect a correct decision, or a type II error. Effect size and sample size are the key factors that determine type II error rate, as will be discussed in [Chapter 20](#).

References

1. Schwartz M, Polgar S. *Statistics for Evidence-Based Health Care* . Croydon, Australia: Tertiary Press; 2003.
2. Coates S.J, Steed L, Price J. *SPSS: Analysis Without Anguish* . Sydney: John Wiley & Sons Australia Ltd.; 2003.

Effect Size and the Interpretation of Evidence

Introduction,
Effect Size,
 Effect Size with Continuous Data,
 Effect Size with Discontinuous Data,
Confidence Intervals and Effect Size,
 Calculating Confidence Intervals,
 Confidence Intervals and Statistical
 Significance,
How To Optimize Statistical Inferences,
How To Interpret Null (Nonsignificant) Results,
Statistical Power Analysis,
Clinical Decision Making,
Overview of Conducting Data Analysis,
 Descriptive Statistics,
 Inferential Statistics,
Summary,

Introduction

When one refers to the *significance* of a research study this may actually be a reference to two quite different types of significance. These are *statistical significance* and *clinical significance*, which refer to two quite different concepts. Statistical significance is concerned with whether the study's results may have arisen because of random sampling errors or they reflect 'real' population trends (i.e. the study sample results accurately reflect those that would be obtained by a study of the full population). On the other hand, clinical significance is concerned with whether the results will matter to the patients (i.e. represent a valuable advance in treatment outcomes or clinical efficacy and effectiveness). It is entirely conceivable that a study may have a statistically significant result but that the magnitude of the improvements is clinically small and not particularly beneficial to the patients. That is, the research findings may be statistically significant but not practically or clinically significant. If research findings are not statistically significant, then clinical significance does not come into play, statistical significance must be demonstrated first before the clinical effects size can be known.

So, after having established statistical significance, we must also establish the clinical or practical significance of our results once we have demonstrated statistical significance.

Two of the world's most eminent statisticians have had pertinent observations to make about the issue of clinical and statistical significance. Gene V Glass the statistician who invented meta-analysis noted:

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them. ¹

Jacob Cohen (1990), who pioneered power analysis noted:

The primary product of a research inquiry is one or more measures of effect size, not P values. ²

As Sullivan and Feinn note, the ‘p value is not enough’ on its own to determine whether a research result is clinically significant.³ So how do we do that (i.e. establish whether a result is clinically significant)? The practical or clinical significance of the results is measured by a statistic referred to as the ‘effect size’; the larger the ‘effect size’, the more likely it is that the results will be clinically useful. As we will see in this chapter, effect size is the most relevant statistic for establishing the efficacy of an intervention.

The aims of this chapter are to discuss:

1. Effect size (i.e. how large the relationships or differences observed in the data are).

Table 20.1

Index	Derivation	Interpretation
Cohen's d	Where X_1 is the mean of group 1, X_2 is the mean of group 2 and s is the standard deviation of one or the other group	Small effect = 0.2 Medium effect = 0.5 Large effect = 0.8
Relative risk	Ratio of the probability of an outcome in group 1 versus group 2	Small effect = 2 Medium effect = 3 Large effect = 4
Odds ratio		Small effect = 1.5 Medium effect = 2 Large effect = 3
Pearson's r	Ranges from -1 to + 1	Small effect = ± 0.2 Medium effect = ± 0.5 Large effect = ± 0.8
Coefficient of determination (r^2)	Ranges from 0 to 1, sometimes expressed as a proportion, sometimes as a percent	Small effect = + 0.04 Medium effect = + 0.25 Large effect = + 0.64

2. The relationship between effect size and statistical and clinical significance.
3. The determinants of statistical power.
4. The basic principles of clinical decision making.

Effect Size

Effect size expresses the size of the change or degree of association that can be attributed to a health intervention. The term ‘*effect size*’ is also used more broadly in statistics to refer to the size of the phenomenon under

study. For example, if we were studying gender effects on how long people live, a measure of effect size for longevity attributed to gender could be the difference in life expectancy between males and females. On average, globally this difference is actually around 5.5 years ⁴, which has real implications! In a correlational study, the effect size is the size of the correlation between the selected variables under study (e.g. r^2 as discussed in [Chapter 17](#)). There are many measures or indicators of effect size; the one which is relevant to analysing the results of a study is selected on the basis of the scaling of the outcome or dependent variable. ⁵

[Table 20.1](#) summarizes some common measures of effect size.

Effect Size with Continuous Data

The concept of effect size for continuous data (i.e. an interval or ratio measurement scale) can be illustrated by results from two student research projects supervised by one of the authors.

Study 1: Test–Retest Reliability of a Force Measurement Machine

In this study, the student was concerned with demonstrating the test–retest reliability of a device designed to measure maximum forces being produced by patients' leg muscles under two conditions (flexion and extension). Twenty-one patients took part in the study, and the reliability of the measurement process was tested by taking two readings for the same patients an hour apart. The Pearson correlation was then calculated between the readings obtained from the machine in question during two trials, separated by an hour for each patient. The results are shown in [Table 20.1](#). Both results reached the $p = 0.01$ level of significance.

The student was ecstatic when the computer data analysis program informed that the correlations were statistically significant at the 0.01 level (indicating that there was less than a 1 in 100 chance that the correlations were illusory or actually zero). We were somewhat less ecstatic because, in fact, the results indicated that approximately 69% ($1 - 0.56^2$) and 71% ($1 - 0.54^2$) of the variation was not shared between the measurements of the first and second trial. In other words, the measures were 'all over the place',

despite statistical significance being reached. Thus far from being an endorsement of the measurement process, these results were somewhat of a condemnation. This is a classic example of the need for careful interpretation of effect size in conjunction with statistical significance.

Table 20.2

Pearson Correlations Between Trials 1 and 2

Flexion	Extension
0.56	0.54

Table 20.3

Mean Adl Disability Scores

	Rsi Group	Normals
Mean	33.2	30.4
Standard deviation	1.6	1.2

ADL, Activities of daily living; *RSI*, repetition strain injuries

Study 2: A Comparative Study of Improvement in Two Treatment Groups

The second project was a comparative study of two groups: one group suffering from suspected repetition strain injuries (RSI) induced by frequent computer data entry and a group of ‘normals’. A scale assessing activities of daily living (ADL) was used and yielded a ‘disability’ index of between 0 and 50. There were 60 people in each group. The results are shown in [Table 20.2](#). The appropriate statistic for analysing these data happens to be the independent group’s *t* test, although this is not important for understanding this example. The *t* value for these data was significant at the 0.05 level. Does this finding indicate that the difference is clinically meaningful or

significant? There are two steps in interpreting the clinical significance of the results.

First, we calculate the effect size. For interval or ratio-scaled data the effect size ' d ' is defined as:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

where $\mu_1 - \mu_2$ refers to the difference between the population means and σ refers to the population standard deviation.

Because we rarely have access to population data, we use sample statistics for estimating population differences. The formula becomes:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_1}$$

where $\bar{X}_1 - \bar{X}_2$ is the difference between the sample means and s_1 refers to the standard deviation of the 'normal' or control group. Therefore for this example, substituting into the equation yields:

$$d = \frac{30.4 - 33.2}{1.2} = -2.33$$

In other words, the average ADL score of the people with suspected RSI was 2.33 standard deviations under the mean of the distribution of 'normal' scores. The meaning of d can be interpreted by using standardized scores. The greater the value of d , the greater the standardized difference between the means and therefore the larger the effect size. The inventor of the d effect size statistic, Jacob Cohen, has argued that $d = 0.2$ can be considered

to be a ‘small’ effect size, 0.5 can be considered to represent a ‘medium’ effect size and ≥ 0.8 can be considered a ‘large’ effect size. These are not hard-and-fast rules—they are rules of thumb.

In terms of considering the clinical implications of the evidence we have just presented, it might be that the difference of 2.8 units in the ADL scores is important and clinically meaningful. However, if one inspects the means, the differences are slight, notwithstanding the statistical significance of the results. This example further illustrates the problems of

interpretation that may arise from focusing on the level of statistical significance and not on the effect sizes shown by the data.

When we say that the findings are clinically or practically significant we mean that the effect is sufficiently large to influence clinical practice (i.e. result in changes in clinical guidelines and clinical decisions concerning what treatments to apply based on the research evidence). It is the health workers rather than statisticians who need to set the standards for each health and illness determinant or treatment outcome. After all, even relatively small changes can be of enormous value in the prevention and treatment of illnesses. There are many statistics currently in use for determining effect size. The selection, calculation and interpretation of various measures of effect are beyond this introductory book, but interested readers can refer to Sackett et al.⁶

However, it is not simply a matter of a clinician deciding whether in their personal opinion an incremental gain in clinical efficacy is ‘worth it’.⁵ Every treatment alternative has a cost associated with it. Analysis of health costs is best dealt with by performing an economic evaluation. Such an evaluation calculates the costs of providing a health care intervention in comparison to its benefits. An economic analysis will usually consider the costs and benefits of conventional or usual care compared with the costs and benefits of a new intervention. Inclusion of consideration of costs and benefits enables consideration of efficiency and hence whether the interventions are the best use of scarce health care resources. Actually, although the convention has been to consider effect size simply in terms of the standard deviation of the difference, a better analysis is to consider the costs and benefits of the comparative interventions. So, although an intervention may only deliver a ‘small’ effect-size increment, if the costs

associated with the illness are very high, this modest effect size may indeed be ‘worth it’.

Effect Size with Discontinuous Data

For a randomized controlled trial (RCT), cohort or case–control study with two groups, effect sizes correspond to the size of the difference between the two groups. For dichotomous data, measures of effect size include the odds ratio (OR), the absolute risk reduction (ARR) and the relative risk reduction (RRR).

OR compares the odds of the occurrence of an event for one group with the odds of the event for another group. For an RCT, the OR compares the odds of an event for the intervention group with the odds for the control group. For example, the odds for each group of having type 2 diabetes, in the hypothetical RCT of an exercise program for obese men. So, using these numbers, the OR is:

$$OR = \frac{0.25}{0.67} = 0.37$$

So, this means that the odds of developing type 2 diabetes in the intervention group are about one-third of those for the control group. An OR of this size could well be interpreted as evidence for the efficacy of the exercise program.

ARR is a simple and useful measure of effect size. It is calculated by subtracting the percentage of people in the control group who experience the event, from the percentage of people in the intervention group who experience the event. The percentage values for each group for the hypothetical RCT of the exercise program for obese men were given previously. Using these numbers, the ARR is:

$$ARR = 20 - 40 = -20$$

This means that the risk for type 2 diabetes was reduced by 20% in the intervention (exercise) group.

RRR is calculated by dividing the ARR by the value corresponding to the percentage of people in the control group who experience the event. For the hypothetical RCT of the exercise program for obese men, the ARR and the percentage of the control group who experience the event (develop type 2 diabetes) were given earlier. Using these numbers from the example, the RRR is:

$$RRR = \frac{20 - 40}{40} = -0.5 \text{ OR}$$

In other words, the risk for developing type 2 diabetes is reduced by 50% in the intervention (exercise) group, relative to the control group. This finding also indicates that the hypothetical treatment was effective; however, we will need to demonstrate the statistical significance of the results.

Confidence Intervals and Effect Size

The calculation of confidence intervals (CIs) is an essential step in analysing and reporting effect sizes. Effect sizes and associated CIs constitute the most important research findings for conducting evidence-based practice. ⁶

Calculating Confidence Intervals

Calculated from the sample data, the CI is represented by a range of values which contain, at a given level of probability, the true population parameter (see [Ch. 18](#)). For example, the 95% CI for an OR will include, at a

probability of $p = 0.95$, the actual OR for the population. There is only a small probability ($p \leq 0.05$) that the population OR will fall outside the lower or upper limits of the 95% CI.

For example, consider the hypothetical RCT discussed earlier in the chapter. The OR for the sample was 0.37. Can we infer in general that the exercise program reduces the odds of type 2 diabetes by *exactly* this number? Of course not; rather the researcher will need to calculate and report the 95% CI. Say that the researchers calculated and reported these results:

$$OR = 0.37 (0.32 - 0.42)$$

These results would indicate that the odds of type 2 diabetes in the population could be reduced to 0.32 (lower limit) or 0.42 (upper limit) by the exercise program. That is, we are not absolutely confident in the actual population effect size (OR), but we can be 95% confident that it is between 0.32 and 0.42. Similarly, CIs (usually but not always 95% CI) are reported for other measures of effect size including RRR, Cohen's d , or r^2 .

The calculation of CIs for effect sizes is beyond the scope of this introductory text. However, interested readers may consult Hicks ⁶ for details on conducting these calculations. In general, the calculated width of a confidence interval depends on the degree of variability of the data and the sample size. The larger the sample size, the narrower the CI. Large sample sizes enable us to collect more information than small samples and then enable us to make precise estimates of the true effect size for the population.

Confidence Intervals and Statistical Significance

As for other statistics, we can calculate the probability of an effect size being just due to chance or an actual effect. An effect size, such as 'OR' or

‘*d*’ is reported as statistically significant if the probability of it occurring by chance is <0.05 ($p \leq 0.05$). There is a close relationship between the CI for an effect size and statistical significance as illustrated by the results of a hypothetical RCT (Fig. 20.1).

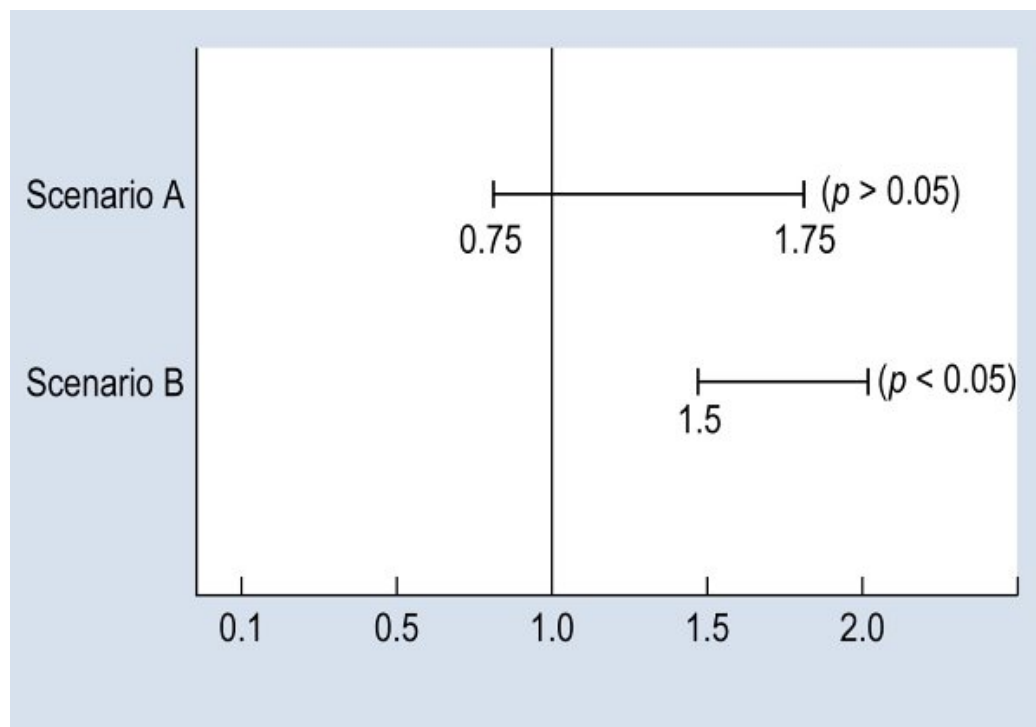


FIG. 20.1 Decision criterion: risk of type II error (miss).

Fig. 20.1 illustrates the results of a study where the outcome measure was on a nominal scale and the results were analysed using OR as the measure of effect size. For ‘scenario A’, $OR = 1.25$ with a lower limit of 0.75 and upper limit of 1.75. This range includes $OR = 1.0$, indicating that the true population value of no effect is included in the range of the CI. These findings indicate that we would retain H_0 : $OR = 1.0$ at a 0.05 level of significance.

For ‘scenario B’, the results were $OR = 2.0$ with a lower limit equal to 1.5 and an upper limit equal to 2.5. Here the $OR = 1.0$ indicating no *effect* is not included in the 95% CI. Therefore we can reject the null hypothesis and accept the alternative H_A :

$$OR \neq 1.0 (p < 0.05)$$

These examples apply to other measures of effect size linking 95% CI and statistical significance. If the confidence interval does not include the null value then the results are deemed to be statistically significant. Effect sizes with 95% CI are the key statistics for determining the efficacy of interventions in the context of conducting evidence-based health care.

How to Optimize Statistical Inferences

It should be evident from the previous discussion that statistical decision making can result in incorrect decisions. There are two main types of inferential error: type I and type II. A type I error occurs when we mistakenly reject H_0 ; that is, when we claim that our experimental hypothesis was supported when it is, in fact, false. The probability of a type I error occurring is less than or equal to α . For instance, if we set $\alpha = 0.01$, then the probability of making a type I error is less than or equal to 0.01; the chances are $\leq 1/100$ that our decision in rejecting H_0 was mistaken. The smaller the α , the less the chance of making a type I error. We can set α as low as possible, but by convention it must be ≤ 0.05 .

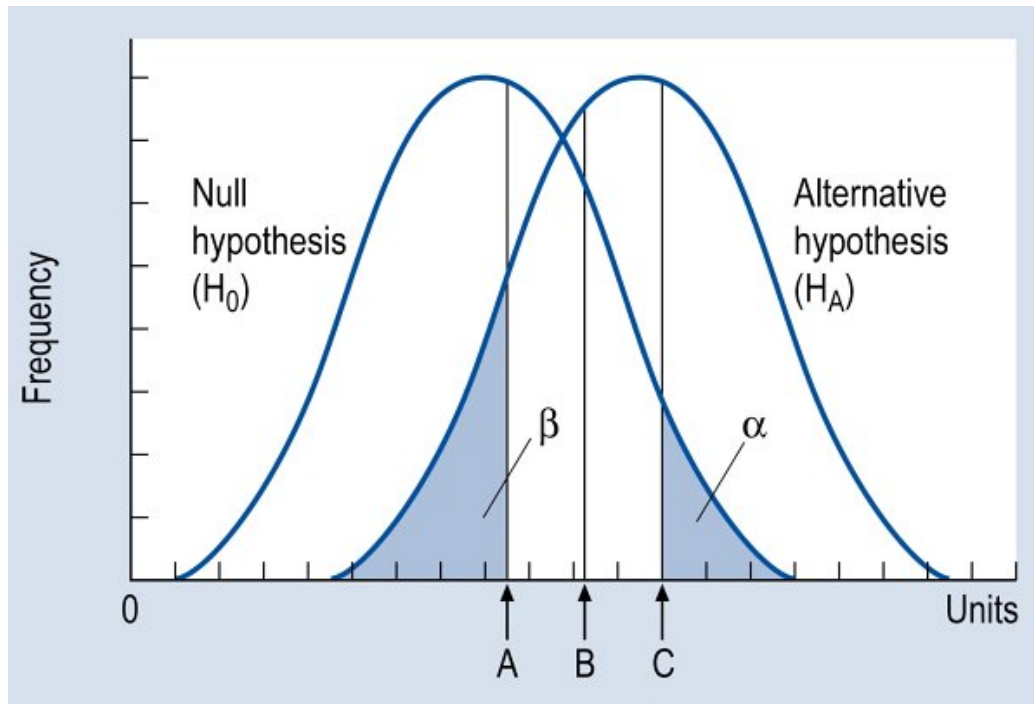


FIG. 20.2 Decision criterion: risk of type I error (false alarm).

A type II error occurs when we mistakenly retain H_0 ; the probability of a type II error is denoted by β (beta). If n is large we are more likely to reject H_0 and conclude that the results are statistically significant.

Type I errors represent a ‘false alarm’ and type II errors represent a ‘miss’. [Fig. 20.2](#) illustrates that if we reject H_0 , we are making either a correct decision or a type I error. If we retain H_0 , we are making either a correct decision or a type II error. Although we cannot, in principle, eliminate type I and type II errors from scientific decision making, we can take steps to minimize their occurrence.

We minimize the occurrence of type I error by setting an acceptable level for α . In scientific research, editors of most scientific journals require that α should be set at ≤ 0.05 . This convention helps to reduce false alarms to a rate of $< 1/20$. Replication of the findings by other independent investigators provides important evidence that the original decision to reject H_0 was correct. How do we minimize the probability of type II error?

1. Increase the sample size, n .

2. Reduce the variability of measurements, either by increasing accuracy or by using samples that are not highly variable for the measurement producing the data.
3. Use a directional H_A , on the basis of previous evidence about the nature of the effect.
4. Set a less demanding α , type I error rate. There is a relationship between α and β , such that the smaller α , the greater β . This relationship is illustrated in [Fig. 20.2](#), which shows that, as α decreases, β increases. Inevitably, as we decrease the type I error rate, we increase the probability of type II error. This is the reason why we do not normally set α lower than $p = 0.01$. Although a significance level such as $\alpha = 0.001$ would reduce ‘false alarms’, it would also increase the probability of a ‘miss’.

Table 20.4

Statistical Decision Outcomes

Reality	Decision: Effect	Decision: No Effect
Effect	Correct	Type II error (miss)
No effect	Type I error (false alarm)	Correct

5. Finally, a large effect size reduces the probability of making a type II error.

How to Interpret Null (Nonsignificant) Results

As we discussed in previous chapters, the researcher will sometimes analyse data using a statistical method and the results show no statistically significant relationships or differences. In other words, the researcher cannot reject the null hypothesis. There are several reasons why the researcher may obtain a null result.

1. The trend or difference that the researcher originally hypothesized is incorrect (i.e. there is no effect).
2. The sample included in the analysis is unrepresentative, so the effect does not show up in the study sample, although it actually exists in the wider population (i.e. the data collected are biased and flawed).
3. There are insufficient cases in the sample to detect the trend; this is especially a problem if the trends are subtle (i.e. the effect size is small; the study sample is not big enough). This is a statistical power consideration. It is essential that null results are considered in the context of appropriate statistical power considerations.
4. The measurements chosen have very high or inherent random variability (i.e. the measurement tools used are not good enough).

Therefore if the researcher obtains a null result, one or more of these explanations may apply. There are, however, actions that can be taken to minimize the chance of missing real effects. To understand these actions, it is necessary again to examine the table illustrating the possible outcomes of a statistical decision (as shown in [Table 20.4](#)).

[Table 20.4](#) shows the four possible decision outcomes just described. On the basis of the statistical evidence you may:

1. Correctly conclude there is an effect when there really is an effect.
2. Decide that there is an effect when there really is not (this is a ‘false alarm’ or type I error).
3. Decide that there is not an effect when there really is (this is a ‘miss’ or type II error).
4. Correctly decide there is not an effect when indeed there is not.

Statistical Power Analysis

The statistical power of a statistical analysis is defined as $1 - \text{probability of a miss (type II error or } \beta \text{)}$.

$$\text{Power} = 1 - \beta$$

That is, the lower the probability of making a type II error, the more likely that we are to make a correct decision. For example, if the power of a particular analysis is 0.95 for a given effect size we will correctly detect the existence of the effect 95 times out of 100. Power is an important concept in the interpretation of null results. For example, if a researcher compared the improvements of two groups of only five patients under different treatment circumstances, the power of the analysis would almost certainly be low, say 0.1. Thus 9 times out of 10, even with an effect really present, the researcher would be unable to detect it. Contemporary health research requires that the power for a study should be at least 0.8. What this means is that we should be 80% certain that we will identify an effect if it is actually true for the population.

It is essential to be careful in the interpretation of null results where they are used to demonstrate a lack of superiority of one treatment method over another, especially when there is a low number of cases. This may be purely a function of low statistical power rather than the equivalence of the two treatments. Unfortunately the calculation of statistical power is complicated and beyond the scope of this text. However, there are technical texts, such as Cohen ⁷ and Ellis, ⁸ that are available to look up the power of various analyses. There are also statistical programs that perform the same function. The best defence against low statistical power is a good-sized sample. Before quantitative research projects are approved by funding bodies or ethics committees, there is the requirement that sufficient data will be collected to identify real effects.

Effect size is also an important consideration for identifying both statistical and clinical significance. We are more likely to detect a significant pattern or trend in our sample data when a factor has a strong influence on health or illness outcomes. A very powerful treatment such as the use of antibiotics for bacterial infections could be demonstrated even in a small sample.

Table 20.5 shows the association between effect size and sample size for determining statistical and clinical significance for the results of research

and evaluation projects. The most useful results for clear decision making occur when both effect size and sample size are large. Where the effect size is large, but the results are not statistically significant, it might be useful to replicate the study with a larger sample size. Unfortunately, in real research it might be difficult to obtain a large sample and the effect size, we discover, might be disappointingly small. It is for this reason that researchers make the best use of previous research and, if possible, complete pilot studies. The evidence from previous research and the results of pilot studies enable us to conduct power analyses for estimating the minimum sample size for detecting an effect if it is really there.

Table 20.5

Effect Size	Sample size	
	Large	Small
High	Both statistical and clinical significance are likely to be demonstrated	Statistical significance might not be demonstrated, but clinical significance would be indicated
Low	Statistical significance would be likely, but the results might not indicate clinically applicable outcomes	Neither statistical nor clinical significance is likely. Statistically significant results might result in a type I error

Table 20.6

Clinical Significance	Statistical significance	
	Yes	No
Yes	Clear: strong evidence for treatment effect	Inconclusive: need for further research (e.g. with larger samples)
No	Inconclusive: suggests findings might not be meaningful-need further research	Clear: strong evidence for lack of a treatment effect

Acknowledgement: We are grateful to Dr. Paul O'Halloran (La Trobe University) for this table.

Table 20.6 shows how evidence for statistical and clinical significance can be combined to interpret the findings of a study. A clear positive outcome is when there is strong evidence for both clinical and statistical significance. In this case we are confident that the information obtained is clinically useful and generalizable to the population. Another clear finding, provided that the sample size was adequate, is the lack of clear treatment effect. Such negative findings can be very useful in eliminating false hypotheses or ineffectual treatments.

Clinical Decision Making

The decisions confronting a clinician making a diagnosis on the basis of uncertain information are similar to the scientist's hypothesis-testing procedure. As a hypothetical example, assume that

a clinician wishes to decide whether a patient has heart disease on the basis of the cholesterol concentration in a sample of the patient's blood. Previous research of patients with heart disease and 'normals' has shown that, indeed, heart patients tend to have a higher level of cholesterol than 'normals'.

When the frequency distributions of cholesterol concentrations of a large group of heart patients and a group of 'normals' are graphed, they have the same shape as shown in [Fig. 20.2](#). You will note that, if patients present with a cholesterol concentration between 1.0 and 2.4 mg/mL, it is not possible to determine with complete certainty whether they are normal or have heart disease, because of the overlap of the normal and heart disease groups in the cholesterol distribution. Therefore the clinician, like the scientist, has to make a decision under uncertainty: to diagnose pathology (i.e. reject the null hypothesis) or normality (i.e. retain the null hypothesis). The clinician risks the same errors as the scientist, as shown in [Table 20.7](#).

The relative frequency of the type of errors made by the clinician can be altered by moving the point above which the clinician will decide that pathology is indicated (i.e. the decision criterion). For example, if clinicians did not bother their colleagues or patients with false alarms

(type I errors), they might shift the decision criterion to 2.5 mg/mL ([Fig. 20.2](#)). Any patient presenting with a cholesterol level <2.5 mg/mL would be considered normal. In this particular case, with a decision point of 2.5 mg/mL, no 'false alarms' would occur. However, a huge number of people with real clinical pathology would be missed (type II errors). If clinicians value the sanctity of human life (and their bank balance after a successful malpractice suit) they will probably adjust the decision criterion to the point shown in [Fig. 20.3](#). In this case there would be no misses but lots of false alarms.

Table 20.7

Clinical Decision Outcomes

Reality	Decision: Pathology	Decision: No Pathology
No pathology	Type I error (false alarm)	Correct decision
Pathology	Correct decision	Type II error (miss)

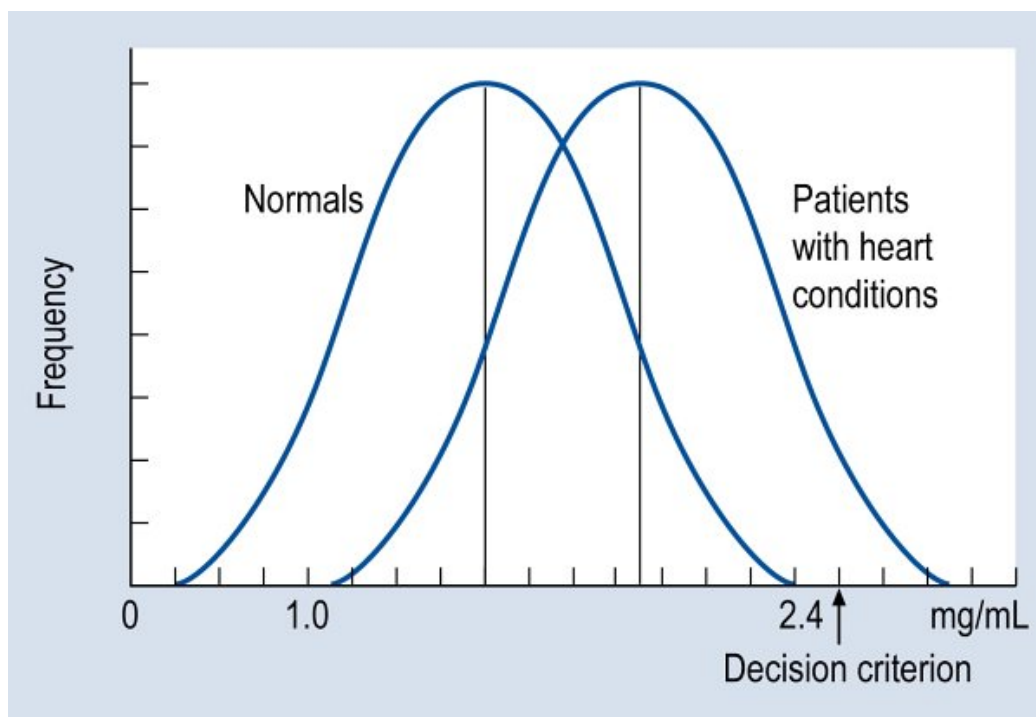


FIG. 20.3 Decision criterion: risk of type II error (miss).

Thus most clinicians are rewarded for adopting a conservative decision rule, where misses are minimized, by receiving lots of false alarms. Unfortunately, this generates a lot of useless, expensive and sometimes even dangerous clinical interventions.

Overview of Conducting Data Analysis

Let's examine a hypothetical example for illustrating how the use of descriptive and inferential statistics is applied when analysing quantitative data. There is evidence that exposure to heavy metals, such as lead, may result in changes in the cognitive abilities of young adults. A researcher selects the design referred to as a *cohort* study (see [Ch. 8](#)), for guiding data collection to answer the research question 'How does significant exposure to lead influence the IQ in young people between 18 and 20 years of age?' Assume that a sample of 100 volunteers between 18 and 20 years of age are identified as having high serum lead levels. This group is tested and compared on IQ with a control group matched on age, gender and socioeconomic background.

The research hypothesis is therefore:

The group of 16- to 19-year-olds exposed to lead will have lower IQs on the Wechsler Adult Intelligence scale than a matched group with no exposure to lead.

Say that the statistical package SPSS is used to analyse the data.

Descriptive Statistics

Having entered the data, the first step is to calculate the relevant descriptive statistics. Given that the variable, IQ, is an interval scale, the appropriate statistics for representing the results are the mean and the standard deviation. *Graphing* the frequency distribution would help visualize the difference between the IQs of the exposed group and the nonexposed (control) group. Say that these results are calculated:

Exposed group $\bar{X}_e = 90$, $S_e = 15$, $n_e = 100$

Non exposed comparison group $\bar{X}_c = 105$,
 $S_c = 15$, $n_c = 100$

$\bar{x}$ is the mean, S the standard deviation and n the sample size for each of the exposed and nonexposed groups. Clearly, there is a 15-point difference between the two groups. This would be a large difference; however, we need to establish if it was due to random error/chance.

Inferential Statistics

To establish the probability of the null hypothesis H_0 , $\mu_e = \mu_c$, we would select the independent group's t test. The hypothetical results would be $t = 7.1$, $df = 198$, $p < 0.001$. Therefore we reject the null hypothesis and say that our results are statistically significant at the 0.001 level. In other words, the findings can be generalized to the population; that is, we infer that young people exposed to lead have lower IQs than those who are not exposed. The effect size enables us to estimate the size of the effect. In this case there was a large 15-point difference in the IQs. Using Cohen's d to calculate the standardized effect size results in $d = 1.0$ (0.8–1.2); 0.8 and 1.2 are the approximate lower and upper limits for the 95% CI. This is a large effect size indicating that the exposure to lead is having serious effect on the cognitive development of young people. Of course, this is just a hypothetical example, but such a large effect size in real life would be a cause for concern that is of practical significance.

Summary

In the interpretation of a statistical test, the researcher calculates the statistical value and then compares this value against the appropriate table to obtain the probability level. If the probability is below a certain value (0.05 is a commonly chosen value), the researcher has established the statistical significance of the analysis in question. The researcher must then interpret the implications of the results by determining the actual size of the effects observed. If these are small, the results may be statistically

significant but clinically unimportant. Statistical significance does not imply clinical importance.

A null result (indicating no effects) must be carefully interpreted. It is possible that the researcher has missed an effect because of its small size and/or insufficient cases in the analysis. The statistical analysis measures the chance of correctly detecting a real effect of a given size.

Thus a null result may be a function of low statistical power, rather than there being no real effect. It is necessary to carry out a statistical power analysis *before* undertaking a research project.

There are several criteria, beyond statistical significance, which need to be considered before making decisions concerning the clinical relevance of investigations. The most important criterion is a large and consistent effect size. In addition to effect size, the determination of clinical or practical significance is influenced by negative side effects of treatments and economic limitations concerning the administration of health care in a given community.

References

1. Glass G.V. Primary, secondary, and meta-analysis of research. *Educ Res* . 1976;5(10):3–8. doi: [10.3102/0013189X005010003](https://doi.org/10.3102/0013189X005010003).
2. Cohen J. Things I have learned (so far). *Am Psychol* . 1990;45(12):1304–1312.
3. Sullivan G.M, Feinn R. Using effect size—or why the *P* value is not enough. *J Grad Med Educ* . 2012;4(3):279–282. doi: [10.4300/JGME-D-12-00156.1](https://doi.org/10.4300/JGME-D-12-00156.1).
4. Raftery J, Young A, Stanton L, et al. Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme. *Health Technol Assess* . 2015;19(11):1–138. doi: [10.3310/hta19110](https://doi.org/10.3310/hta19110).

5. Sackett D.L, Strauss S.E, Richardson W.S, Rosenberg W, Haynes R.B. *Evidence-Based Medicine: How to Practice and Teach EBM* . 2nd ed. Edinburgh: Churchill Livingstone; 2000.
6. Hicks C.M. Applied project design and analysis. In: Hicks C.M, ed. *Research Methods for Clinical Therapists: Applied Project Design and Analysis* . 5th ed. London: Elsevier Health Sciences; 2009.
7. Cohen J. *Statistical Power Analysis for the Behavioural Sciences* . 2nd ed. London: Routledge; 1988.
8. Ellis P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* . Cambridge: Cambridge University Press; 2010.

Qualitative Data Analysis

Introduction,
Understanding Meaning in Everyday Life,
Management and Organization of Data,
Coding Qualitative Data,
 Predetermined Coding,
 Coding and Thematic Analysis,
Content Analysis,
 Using Computer Programs to Evaluate
 Qualitative Data,
Thematic Analysis, 'Verstehen' and Grounded Theory,
Interpretation and Social Context,
The Accuracy of Qualitative Data Analysis,
Summary,

Introduction

As we have discussed previously, the aim of qualitative research is to provide evidence for understanding experiences of health and illness from the perspectives of patients and research participants. Qualitative data are

collected through techniques such as in-depth interviews, focus groups and participant observation. In this chapter, we will examine qualitative data analysis. Qualitative data analysis refers to the processes by which researchers organize the information collected and analyse the meanings of what was said and done by the participants.

When conducting qualitative data analysis we bring our values, experiences and theories to analysing and constructing the meaning of what our respondents are telling us about their experiences of health, illness and health care in the context of their social settings.

Before we look at the details of qualitative analysis, we examine an overall model of the process ([Fig. 21.1](#)).

The specific aims of this chapter are to:

1. Describe the process of interpreting qualitative research data.
2. Describe the basic procedures involved in conducting content analysis, thematic analysis and semiotic analysis.
3. Discuss the comparative advantages and disadvantages of using different types of qualitative analyses.
4. Explain basic strategies for ensuring the accuracy of interpretations.

Understanding Meaning in Everyday Life

Understanding people involves discovering their beliefs, desires, intentions. We do this by listening to them and observing what they say and do, taking into account the social settings in which these actions occur. In many ways, qualitative data analysis is the formalized and principle equivalent of understanding and forming empathic relationships with people.

For instance, one person says to another: ‘Would you like to come in for a coffee?’ What are the intentions of the speaker? Does the speaker simply want to prepare the coffee and consume it in silence? Or should we look for ‘hidden’ or ‘latent’ meanings to understand the speaker’s true intentions? Consider these two everyday scenarios:

1. You have been given a lift by a workmate who had to go out of his way to drive you home. Although you are tired, it seems the right thing to offer the driver refreshments. However, your intention is to

- be polite and acknowledge the colleague's effort; in fact, you are hoping that the invitation will be refused. The 'hidden message' here is 'Thank you and goodbye!' The worst-case outcome is that the colleague is too insensitive to read your intentions and stays around gossiping until midnight. Bad luck!
2. The British film 'Brassed Off' (1996) has a scene in which a young woman is escorted home after a date by a young man. A dialogue was (approximately) as follows:

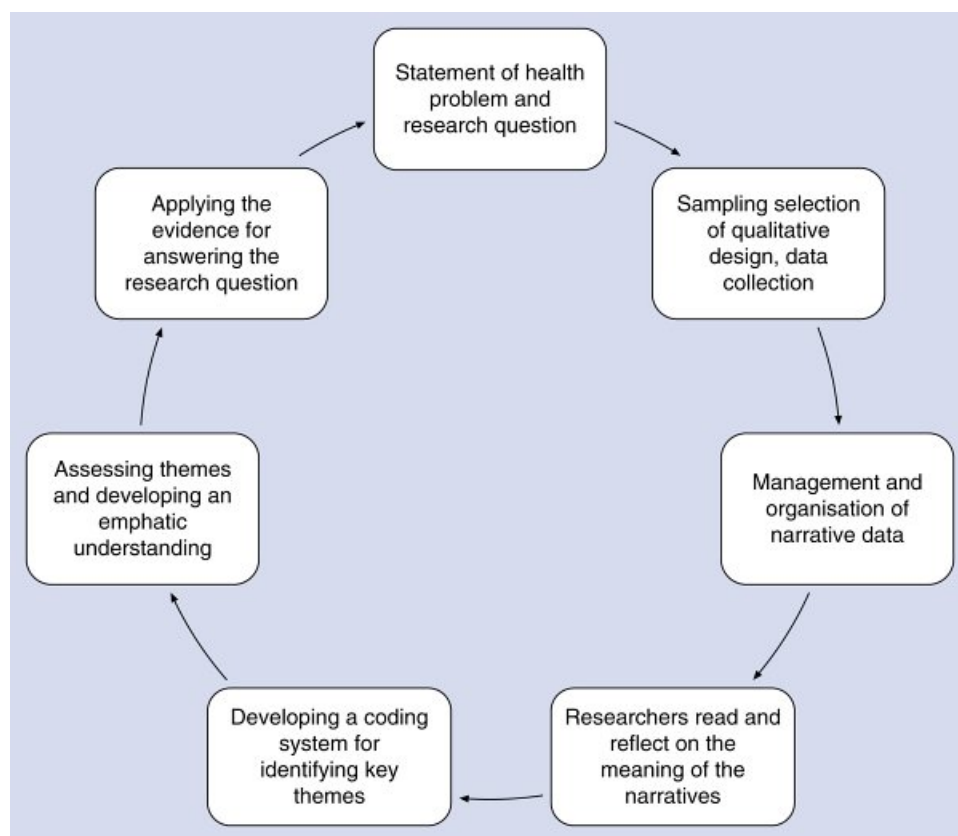


FIG. 21.1 The process of qualitative data analysis (after Creswell JW, Poht CN. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. 4th ed. Thousand Oaks, CA: Sage; 2018.).

She: Come up for a cup of coffee.

He: I don't drink coffee.

She: That's alright; I don't have any.

The dialogue shows the nuances in the everyday use of language. Just as we sometimes misunderstand meanings and intentions in everyday life, we can also misinterpret the data produced by qualitative data collection. To avoid error, we need to cross-check the accuracy of our interpretation.

In the sections that follow, we will examine the stages of the qualitative research process in more detail.

Management and Organization of Data

At the conclusion of the data collection process, qualitative researchers are confronted by transcripts which record the narratives expressed by the participants. The first step of analysis is to thoroughly read the transcripts, that is to ‘immerse’ oneself in the text. We need to understand, at a subjective personal level, what the research participants are telling us about their lives. Of course, this information is likely to be personal, and there are very strong ethical requirements to ensure confidentiality. At the conclusion of this ‘immersion’, the qualitative researcher will have constructed a plausible theory about the meaning conveyed by the texts.

Coding Qualitative Data

Once we have acquired an overall understanding of the narratives the next step is to code the text. To do this the researchers develop a coding system which is used to organize the data (what was said) into categories based on similarities of words, concepts and themes. There are two main approaches to coding: *predetermined* and *emerging with thematic analysis*.

Predetermined Coding

Predetermined coding uses predetermined categories to organize and analyse the transcripts.

For example, you might do a survey to study clients’ experiences of a rehabilitation program at your workplace. Say that you conducted 20 in-depth interviews and produced a 100-page transcript representing what people said in these interviews. Considering your research aims, you might code the statements into three categories:

1. Satisfaction with the rehabilitation program.
2. Dissatisfaction with the rehabilitation program.
3. Neutral statements.

At the simplest level, analysing the first two categories would enable us to understand the reasons why the clients found the rehabilitation programs to be satisfactory or unsatisfactory. This information could be useful for improving the program. In most studies the coding system would be more elaborate.

Coding and Thematic Analysis

An alternative approach to using predetermined codes is to develop a coding system that identifies common themes as they emerge from the text. Different qualitative researchers advocate different approaches to coding, but it typically involves the following steps. The researchers first study their transcripts and develop a close familiarity with them. During this process, all the concepts, themes and ideas are noted to form major categories. Often, the researcher will then attach a label and/or number to each category and record their positions in the transcript. Coding is an iterative process (we retrace our steps), with the researcher coding and recoding as the interpretation progresses. The researchers, having developed the codes and coded the transcripts, then attempt to interpret their meanings in the social context of the participants.

Coding is rather like cutting and pasting together similar things; it is an essential process for identifying segments of a text which convey similar meanings. There are a number of units of analysis that can be used to examine meanings in the text. These units reflect the way language is structured and used for communication and can be the bases for coding.

Words: can be identified in the way they have been used, whether descriptively or as expletives.

Concepts: these are words or phrases that convey complex ideas or meanings such as 'disability' or 'stigma'.

Sentences: are grammatically correct sequences of words which convey meaning in a language. Sentences

conveying similar meanings can be coded and grouped together.

Themes: are ideas that are represented in a pattern or with a certain degree of frequency within a text. They are made up of a number of concepts which represent experiences and meanings the individuals attach to them. For example, references to hostile intentions in other people that often emerge in the discourses of people having been diagnosed with paranoid schizophrenia.

Consider the passage which appeared in Polgar and Swerissen (not in print):

*Difficulties with moving, walking and shifting weight can be part of the everyday experience. The inability to move when others are moving, to join in group activities that involve some mobility, and the inability to participate in sport are characteristic of the problems faced by individuals with disabilities. When identity can be gained from joining in physical group activities, people with disabilities can face disadvantage. If it is participation or just companionship that is desired, physical activities do not provide opportunities for some of those with disabilities and can lead to a sense of separation.*¹

We can identify in this text the words, phrases and sentences that contain synonyms (words that mean the same thing). These terms can be coded together and are therefore grouped. *Moving, walking, shifting weight*: these are aspects of the same concept, as representing aspects of physical activity not achievable by those who are severely disabled. In this text, these terms have been used several times. *Move when others are moving; join in group activities that involve some mobility; participate in sport*: these are phrases conveying synonyms of physical disadvantage that can reduce the ability of some people with severe disabilities to participate. Also, there are sentences about physical activities with others. *When identity can be gained from joining in physical group activities, people with disabilities can face disadvantage. If it is participation or just companionship that is desired, physical activities do not provide opportunities for some of those with*

disabilities and can lead to a sense of separation. These sentences convey similar themes about the disadvantage faced by some people with disabilities through their inability to participate in physical activity. *Disadvantage* and *separation* are two emotive terms signifying problems faced by some people with disabilities. These terms can be coded together, demonstrating the way concepts can be used to code qualitative data.

Content Analysis

Content analysis enables the identification of units of meaning occurring in a text. Content analysis is a technique which can be seen as a blending of quantitative and qualitative methods. The recognition and coding of meaning are qualitative, whereas the counting of the meaningful ‘chunks’ is quantitative. The ‘meaningful chunks’ can be words, sentences or paragraphs: that is, the units of language that were coded by the researchers from the narratives and dialogues.

To illustrate content analysis, consider the following unpublished study. One of the authors was interested in how leading newspapers were representing the use of stem cells in medical research. Two research questions were asked:

1. How extensive was the newspaper coverage of the medical use of stem cells?
2. What was the attitude of the newspapers (positive, negative or neutral) to the use of stem cells?

In relation to question 1, the data were collected by identifying relevant newspaper articles published on the topic and measuring the length of the columns. They were quantified by counting the number of articles published per month across the selected time interval. The data relevant to question 2 were obtained by identifying statements supportive or critical of using stem cells or simply ‘neutral’ descriptions of the nature and possible uses of stem cells. The column lengths for each of the three categories were measured and the percentages devoted to each were graphed across the months. Therefore the content analysis provided evidence for the level of interest and changing attitudes of the media towards the use of stem cells.

This evidence was relevant to understanding the cultural context in which government policy for using stem cells was being formulated. Content analysis is a technique that combines elements of both qualitative and quantitative approaches. We interpret the meaning of the text for developing our coding strategy for organizing or ‘chunking’ the text and then we use statistics to describe the quantities of text devoted to a specific point of view. Content analysis can be used to test hypotheses, for example hypotheses addressing the positive and negative media perspectives at a given point in time on embryonic stem cell research. The statistics used to analyse the data were the same as we discussed for quantitative data.

Using Computer Programs to Evaluate Qualitative Data

The discussion of content analysis provides a good opportunity for raising the issue of computer-assisted data analysis. As the texts are often transcribed using personal computers, the text is available in electronic form. This means that the text can be fed into a software package, such as NVivo to assist with its analysis. For example, say that the data representing the contents of one hundred newspaper articles were transcribed into a software package. We could now introduce our codes and identify the segments of the whole text which use the relevant words/phrases/sentences. The segments can then be retrieved, examined or modified (cut and paste) on screen. Also, various frequency counts can be readily performed using software tools.

A detailed discussion for selecting and using computer packages is beyond the scope of the present book. Interested readers might find Rice ² a useful introduction to selecting and using currently available software packages for expediting and improving qualitative data analysis in general (not only for content analysis). There is an ambivalent attitude among qualitative researchers to computer-assisted data analysis. A key objection has been the distancing of the researcher from the creativity and surprising insights afforded by the more hands-on approaches. Another objection is that meanings of words and sentences sometimes do not follow dictionary definitions but rather have to be understood in the general context. The true

meaning of certain subtle and ambiguous communications can be missed in crude and electronically conducted data analyses.

Thematic Analysis, ‘Verstehen’ and Grounded Theory

Counting and hypothesis testing are not the essence of the qualitative approach. What we are trying to do is to see things from the perspectives of our informants and to explain their actions from their points of view. The German word *verstehen* is often used in phenomenological research to express the notion of ‘putting ourselves in someone else’s shoes’ or attaining a strong empathy with their situation. Empathy with other people might seem quite simple, just something we do as human beings. It is worthwhile remembering, however, that sometimes we misunderstand how people feel or think, even when they are our close friends or family. In the same way, we might misunderstand the points of view of persons who are very different to us in age, gender, education, language and culture. Yet it is essential to understand the points of view of the people to whom we offer health services. So how does ‘verstehen’ arise through qualitative health research?

First, as we described earlier, our data collection must use a technique (in-depth interviews, written materials, focus groups, etc.) that enables our respondents to express their point of view. Second, we can adopt a theoretical framework for explaining our understanding of the respondents’ experiences. The key point, in the context of *grounded theory*, is that our explanations or theories must emerge inductively from the information provided by our informants. The theory is constructed gradually as more evidence is provided by additional informants. Third, the data are often analysed by coding and thematic analysis as we outlined earlier in this chapter.

A theme is a grouping of ideas or meanings which emerge consistently in the text. The themes emerging from the data illuminate the experiences of the informants and enable us to understand their points of view (‘verstehen’). Let us consider an example of thematic analysis.

In a study titled ‘The plight of rural parents caring for adult children with HIV’, McGinn studied the experiences of parents caring for their adult children with AIDS/HIV. ³ In-depth interviews were conducted with eight mothers and two fathers from rural families involved in this task. The interview transcripts were analysed using a thematic analysis/grounded theory approach. ⁴

McGinn ³ extracted three major themes:

1. *Physical and mental problems related to HIV/AIDS.* Here the parents discussed their experiences of their children’s problems and the emotional consequences of physical decline and death. For example:

He would fall over, so I would sit him in the wheelchair. And then from within a week in November he went from not being able to sit in the wheelchair to not getting out of bed. And he went from eating little bits of food along with taking a liquid nutrition to just liquid nutrition... and then he got to where he wouldn’t swallow the liquid nutrition and he subsisted on just water and juices and Pepsi... and then in the end he even refused them: he wouldn’t take anything. ... He just wasted away.

2. *Stigma associated with having AIDS.* Because of the mode of transmission of AIDS and superstitious fears of contracting the condition, many of the parents found themselves socially isolated at such a very difficult time of their lives. For example:

That Sunday, I never will forget. I asked him, ‘Do you want anybody to know?’ And I don’t remember if he said no, but his head ... he almost shook it off. No way did he want anybody to know what the real problem was. But I want you to know that that was a terrible, stressful time. People who came, who normally would be support for me ... weren’t. It was a real traumatic experience.

3. *Health care.* This theme summarizes the difficulties of accessing necessary health services in rural settings. Although there were

serious deficiencies in health services, one mother reported:

As for the hospital, I couldn't have asked for a better hospital. There may have been nurses who refused to work with him, I don't know, but the nurses that did come in were great. ... They even hugged and kissed him goodbye whenever he got well and left. They didn't act like they didn't want to be around him, and I appreciated that. I think that's important.

These three themes enable us to understand and empathize with the parents of these very sick young people. Also, they were the bases for recommending improvements in rural health care which directly address the needs of AIDS sufferers and their families in nonmetropolitan environments.

Also, we must note that McGinn's paper reported the experiences of people in the mid-1990s living in rural Canada.³ With improvements in the treatment and prevention of AIDS and a decrease in the stigma attached to the condition, the experiences of families caring for sufferers have improved. Because of differences and changes in practices and the cultural context, it is always important to note the time and place at which interpretive research was carried out.

Interpretation and Social Context

As we have seen, qualitative data analysis is a systematic way of interpreting texts. There are many areas of study (e.g. history, politics, theology) where the interpretation of texts is an essential part of the research process. What these diverse disciplines have in common with qualitative health research is the recognition that the meaning of language and texts must be interpreted in a cultural context.

An example is *hermeneutics*, which is a method that was originally used to analyse the meaning of religious texts. Consider the meaning of the term 'god'. When the Romans spoke of Augustus Caesar as a 'god', they were referring to him as a hero who was immortal in the history of Rome. The use of the term 'god' by a polytheistic is quite different to meanings in the context of contemporary Judaeo-Christian or Muslim traditions. The meaning of the term must be interpreted in the context of the religious

tradition (polytheistic, monotheistic) and the position of the speaker (believer, nonbeliever).

An important issue in reading texts is that they might have implicit (in addition to explicit) meanings. *Semiotics* is a method of textual interpretation which seeks to uncover the hidden, omitted meanings implicit in a text. To do this, we must adopt a theoretical framework in terms of which we can ‘deconstruct’ a text. The theoretical framework reflects our understanding of the culture within which the text was produced. You may have read the book *Animal Farm* by George Orwell. There are several levels at which one can read this story; for example:

- A fairy tale about the imaginary lives of farm animals where animals have human traits and concerns.
- A morality tale in the style of the classical Greek writer Aesop’s style about how power corrupts and leads to betrayal.
- A critique of Stalinism and a retelling of the bloody history of the Bolshevik revolution and its social consequences in the Soviet Union late 1930 and 1940s.

To identify Orwell’s book as a political critique, one needs to understand the historical/cultural context in which the author worked and lived and the political issues which occupied him. If you read the book you may ask why George Orwell wrote a fairy tale with talking farm animals rather than a direct denouncement of totalitarianism and Stalinist terror. It is a question of how we use language; we use metaphors, parables, hyperboles and so on for expressing ourselves in an interesting, colourful fashion. Semiotics is one of the ways for interpreting the meanings that might be hidden or camouflaged in the original narrative.

A basic principle for semiotic analysis is selecting a theoretical framework in terms of which we can deconstruct the original narrative and identify its hidden, repressed or mystifying elements.

Of course, there is always the possibility that we have misinterpreted the text and misrepresented the intentions of the author. The fact of the matter is that there are no absolute guarantees. It might be that, regardless of his well-known interest in political affairs, George Orwell was simply intending to create a children’s story when he wrote *Animal Farm*.

In the next part of this chapter, we will outline some strategies for ensuring validity and reliability for qualitative research.

The Accuracy of Qualitative Data Analysis

How can we be sure that the themes we identified in a text accurately reflect the actual views of the participants? Also, how do we know that similar themes would emerge from the reports of other people who had similar experiences to our sample?

There are a number of qualitative researchers who ensure that the collection and interpretation of their evidence are carried out in a rigorous fashion. Some of the key methodological criteria for conducting qualitative research are discussed here: ⁵

1. *Data saturation*. This refers to ensuring that we have collected sufficient data from our respondents. Saturation occurs when the themes and ideas emerging from the text become repetitive and we are confident that the inclusion of new participants or further engagement with current participants will not lead to novel themes or interpretations.
2. *Credibility*. Checking if the interpretation of the evidence is judged as accurate by both the research participants and also independent clinicians or scholars. In other words, does your interpretation make sense and if not, why not?
3. *Auditability*. This refers to each of the steps of the research process being clearly described, so that an independent scholar can critique the research process from its beginning to the analysis and interpretation of the data. The auditor confirms or rejects the researcher's methodology.

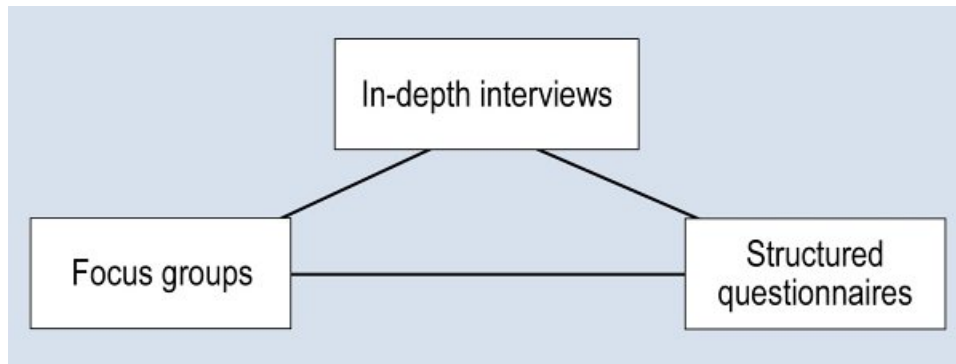


FIG. 21.2 Triangulation using different approaches to data collection.

4. *Triangulation*. This strategy involves the use of multiple independent methods for collecting data and checking if the themes and interpretations emerging from these different methods are consistent and matching. For example, to evaluate client satisfaction with a health service you might use three different data collection strategies (Fig. 21.2). It is useful (but not essential) to use three different data collection methods. The researcher might, if appropriate, use both qualitative (e.g. in-depth interviews) and quantitative (e.g. structured questionnaires) methods for data collection. These are called ‘mixed’ designs.

As you can see, the methodological concerns in qualitative research are parallel to those of quantitative research (i.e. reliability and validity). However, because of the differences in the way the two types of research are conducted, the terminology for describing the methodological principles is somewhat different. There is another difference to keep in mind; some of the facts about the world discovered by physical, biological and behavioural scientists are held as universal truths, even if the laws and theories integrating these facts are open to revision (see Ch. 1). For example, the structure of human DNA as discovered over the past 50 years is assumed to be true for all humans. In contrast, qualitative evidence depends on time, place and socioeconomic circumstances. For instance, the experiences of rural parents caring for adult children with HIV/AIDS is specific to the mid-1990s when the data were collected by McGinn.³ Things have changed; the treatment of people with HIV/AIDS is far more effective and the stigma associated with this illness has become less prejudicial, at least

in some communities. Therefore the experiences of people caring for those with HIV/AIDS 20 years after this research are likely to be different.

Summary

There are different approaches to analysing qualitative data depending on the theoretical framework and data collection strategies adopted by the researchers. However, as we saw in this chapter, there are several common aspects to qualitative data analysis:

- ‘Immersion’ in the data: reading and rereading the texts to develop a sense of what it is that respondents are trying to say.
- Developing a coding system and identifying the themes emerging from the text.
- Using these themes as a basis for insight, empathy (‘verstehen’) with the experiences, emotions and thinking of the respondents.
- Interpreting and theorizing the respondents’ experiences in the context of cultural or historical settings.
- Ensuring the accuracy of the interpretations by cross-checking themes and explanations with other sources of data, other researchers’ interpretations of the data, and the respondents.

References

1. Polgar S, Swerissen H. *Research Methods in Health 2* . Bundoora, Victoria, Australia: La Trobe University; 2000.
2. Liamputtong P. Qualitative data analysis: conceptual and practical considerations. *Health Promot J Aust* . 2009;20(2):133–139.
3. McGinn F. The plight of rural parents caring for adult children with HIV. *Fam Soc* . 1996;77(5):269–278.
4. Miles M.B, Huberman A.M. *Qualitative Data Analysis: A Sourcebook for New Methods* . Beverly Hills,

California: Sage; 1984.

5. Lincoln Y.S, Guba E.G. *Naturalistic Inquiry*
. London: Sage; 1985.

Critical Evaluation of Published Research

Introduction,

Guidelines for the Critical Appraisal of Research Publications,

Critical Evaluation of the Introduction in a Research Paper,

- Adequacy of the Literature Review,
- Clearly Defined Aims or Hypotheses,
- Selection of an Appropriate Research Strategy,
- Selection of Appropriate Variables/Information To Be Collected,

Critical Evaluation of the Methods Section in a Research Paper,

- Research Participants,
- Instruments/Apparatus/Tools,
- Procedure,

Critical Evaluation of the Results Section in a Research Paper,

Critical Evaluation of the Discussion Section in a Research Paper,

Introduction

By the time a research report is published in a peer-reviewed journal, it has been critically reviewed by several experts involved in the editorial process. Usually, changes have been made to the initial draft submitted by the author(s) responding to the reviewers' comments and requests for changes. Nevertheless, even a thorough evaluation procedure doesn't guarantee the validity of the design nor the conclusions presented in a published paper. Ultimately, as a health professional, you must be responsible for judging the validity and relevance of published material for your own clinical activities (which is why you are studying health sciences). Evidence-based practice focuses on the ways in which practitioners can improve their clinical practice by using research evidence. The systematic review processes employed by bodies such as the Cochrane Collaboration are intended to assist clinicians in the selection of interventions that are well proven and safe.

The proper attitude to take with published material, including systematic reviews, is one of hard-nosed scepticism, whatever the status of the publication. This attitude is based on our understanding of the uncertain and provisional nature of scientific and professional knowledge, as outlined in [Chapter 1](#). In addition, health researchers deal with the investigation of complex phenomena, where it is often impossible for ethical reasons to exercise the desired levels of control or to collect crucial information required to arrive at definitive conclusions. The conduct of a valid critique requires that we compare the methods used by the researchers with the rules of evidence in the context of a research project. Critical evaluation identifies the strengths and weaknesses of a research publication, to ensure that patients receive assessment and treatment based on the best available evidence.

The aim of this chapter is to discuss the evaluation of published research. The chapter is organized around the evaluation of the specific sections of research publications.

The specific aims of this chapter are to:

1. Examine the criteria used for the critical evaluation of a research paper.
2. Discuss the implications of identifying problems in design and analysis in a given publication.
3. Outline briefly strategies for summarizing and analysing evidence from a set of papers.
4. Discuss the implications of critical evaluation of research for health care practices.

Guidelines for the Critical Appraisal of Research Publications

Many textbooks on health research methodology or evidence-based health care provide guidelines for the critical appraisal of published research. The Cochrane Collaboration, which is focused on the rigorous evaluation of randomized controlled trial (RCT) research has a detailed manual which contains a very detailed discussion of critical appraisal of RCT studies. The 2011 manual (see <https://training.cochrane.org/handbook>) is currently under major revision and is scheduled for release in 2019. Cochrane's study review manual and protocol are centred on the identification of sources of 'bias' in RCT studies. It categorizes bias into the scheme presented here:

1. **Selection bias.** 'Systematic differences between baseline characteristics of the groups that are compared'. In other words, the treatment and control groups are not equivalent at the commencement of the study.
2. **Performance bias.** 'Systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest'. How the study is conducted (e.g. if the clinicians and/or researchers might let their expectations influence the results if they know which treatments are allocated to which participants). This could also include history and maturation biases (e.g. some participants do unauthorized extra treatment).

3. **Detection bias.** Systematic differences between groups in how outcomes are determined. This might arise from the researchers who are involved in collecting outcome data influencing the results and introducing bias.
4. **Attrition bias.** Systematic differences between groups in withdrawals from a study. In most trials there is some attrition where participants withdraw from the study or it is not possible to collect data from some of them.
5. **Reporting bias.** Systematic differences between reported and unreported findings. The researchers may choose to report only results that are consistent with their predictions/hypotheses and ‘forget’ to report data that does not support their ideas. This is why study registration schemes require a clear statement about what is going to be collected and hence reported so that this form of bias can be detected.

It should be remembered that Cochrane is focused on RCTs and there are many other types of studies for which there are additional risks for bias and which do not include RCT characteristics.

Hicks suggests a list of questions relevant for conducting rigorous appraisals and provides a detailed sample critique of a health research publication.¹ Critical appraisal is an essential step in conducting evidence-based health care. Consequently, textbooks in this area devote chapters to explaining critical evaluations. For example, Dawes et al. provide detailed guidelines for critiquing publications based on different types of designs, including RCTs, case-control and cohort studies and systematic reviews.²

There are also numerous websites hosted by reputable organizations for facilitating the appraisal of published evidence. Examples include:

<https://bestpractice.bmj.com/info/toolkit/ebm-toolbox/critical-appraisal-checklists/>

<https://www.nice.org.uk/process/pmg4/chapter/appendix-f-quality-appraisal-checklist-quantitative-intervention-studies>

<https://www.nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias>

Some groups created critical appraisal schemes which attempt to produce numerical scores to represent methodological quality. For example, the

Downs and Black quality index generates an overall score out of 21 to represent the quality of a publication. Publications with low scores may be excluded from systematic reviews or attributed with poor credibility. Although such quantitative approaches are followed by some reviewers to identify risk of bias (e.g. Soares-Weiser et al. ³ ; see [Ch. 3](#)) it is worth noting that the critical appraisal of published research is not easily qualified. Rather, health researchers or practitioners need to make critical judgements based on their personal knowledge, experience and work contexts.

[Table 22.1](#) summarizes some of the potential problems, and their implications, which might emerge in the context-critical evaluation of an investigation. A point which must be kept in mind is that, even where an investigation is flawed, useful knowledge might be drawn from it. The aim of critical analysis is not to discredit or tear down published work, but to ensure that the reader understands its implications and limitations with respect to theory and practice. In this chapter, we will focus on the constructive, critical evaluation of each stage of the research process as identified in [Chapter 3](#).

Critical Evaluation of the Introduction in a Research Paper

The introduction of a paper essentially reflects the planning of the research. Inadequacies in this section might signal that the research project was erroneously conceived or poorly planned. The issues discussed here are essential for evaluating this section.

Adequacy of the Literature Review

The literature review must be sufficiently complete, including all relevant publications, so that it accurately presents the current state of knowledge in the area. Key papers should not be omitted, particularly when their results could have direct relevance to the research hypotheses or aims. Researchers must be unbiased in presenting evidence that is unfavourable to their personal points of view. This is why we now have systematic review

procedures, such as those utilized by the Cochrane Collaboration, so as to avoid inappropriate and biased exclusion or inclusion of work that supports or challenges a point of view favoured by the researcher or other researchers who hold contrary opinions. Poor review of the literature could lead to needlessly repeating research or making mistakes that could have been avoided if the previous works' findings had been incorporated into formulation of the research design.

Table 22.1

Problems Which Might Be Identified	Possible Implications in a Research Article
1. Inadequate literature review	Misrepresentation of the conceptual basis for the research. A systematic review methodology adhering to PRISMA standards is favoured.
2. Vague research aims or hypotheses	It is preferable that the variables included in the statement of aims are operationally defined (i.e. the ways in which they are to be measured are clearly specified). If the study is an intervention study, whether it is an efficacy or effectiveness (pragmatic) trial should be clearly stated. A PICOT format (Population, Intervention, Comparator, Outcome, Time-frame) for presenting the research questions is preferable.
3. Inappropriate research strategy for the stated aims	The study design must match the study aims. If this is not the case, then research findings might be weakened.
4. Inappropriate variables selected	Measures should be based on recognized measurement protocols and valid high-quality tools that directly relate to the research questions.
5. Inadequate study sampling method	The study sample might be biased and thus the investigation could lack external validity. The target population must be clearly stated.
6. Inadequate sample size	Sample might be biased; statistical analysis might lack power. A power analysis should be provided.
7. Inadequate description of sample	The study sample and the sampling method must be described to the level that it could be easily replicated. Application of findings to specific groups or individuals might be difficult.
8. Measurement instruments lack validity or reliability	Standardized tools with known and published performance are strongly favoured. Self-invented tools should be avoided. Findings might be affected by measurement errors.
9. Inadequate design	The research must be structured so as to avoid the standard biases. If the study is an intervention study, the use of allocation methods and how allocation concealment was implemented must be stated.
10. Lack of adequate control and comparison groups	If the study is an intervention study, the trial purpose (e.g. superiority, equivalence, etc.) must be stated and the design of the comparison group be directly related to that purpose.
11. Biased participant assignment	Allocation sequence and methods and allocation concealment methods must be clearly described and appropriately implemented to avoid allocation biases.
12. Variations or lack of control of treatment parameters	The methods whereby protocol deviations were minimized need to be clearly stated. These might include treatment manuals, training and monitoring of protocol deviations.
13. Observer bias and expectancy effects not controlled (Rosenthal effects)	The processes whereby observer expectancy effects are controlled must be clearly specified. In qualitative research, protocols must include provision to minimize individual coder biases.
14. Participant expectations not controlled (Hawthorne effects)	A clear analysis of how being measured or observed might affect participant expectations and results needs to be provided along with a description of how they were minimized.
15. Research carried out in inappropriate setting	If the purpose of the trial is an efficacy trial, then the setting must be tightly controlled. If the trial purpose is an effectiveness or pragmatic trial, then the setting must be representative of those experienced by the target population to which the findings will be generalized. Otherwise, the investigation might lack ecological validity.

Table Continued

Problems Which Might Be Identified	Possible Implications in a Research Article
16. Confounding of times at which observations and treatments are carried out	In addition to the choice of high-quality measurement tools, the timing and circumstances of all measures must be specified and any effects of timing and sequencing of measurements must be provided.
17. Ensuring adequate presentation of descriptive statistics	The presentation of results must include standard measures such as mean, standard deviations and confidence intervals in order to readily permit subsequent meta-analysis.
18. Avoiding the use of inappropriate statistics to describe and/or analyse data	The statistical tools employed in the study must meet the required assumptions tests and these should be reported. The agreed-on best analysis methods must be implemented.
19. Avoiding erroneous calculation of statistics	The tools used to calculate the statistics must be stated. Recognized statistical packages should be employed in the calculations.
20. Drawing incorrect inferences from the data analysis (e.g. type II error)	All results must be carefully documented with reference to likelihood of type II inferential errors (i.e. false-negative results). This should be discussed with consideration of the statistical power calculations.
21. Protocol deviations	Investigation might lack external or internal validity if the study procedures do not follow the stated protocols.
22. Overgeneralization of findings	The target population and settings to which the results may be generalized must be clearly stated.
23. Statistical and clinical significance should be clearly and separately discussed.	The effect sizes of key findings should be clearly presented in the context of the power analysis.
24. The findings should be clearly contrasted and compared with relevant previous research findings	The commonalities with and differences from other relevant studies should be clearly discussed.

Clearly Defined Aims or Hypotheses

As stated in [Chapter 4](#), the aims or hypotheses of the research should be clearly stated. If this clarity in expression of the aims is lacking, then the rest of the paper will be compromised. In a quantitative research project, it is usual to see a clear statement of the hypotheses and the research aims. All research, whether qualitative or quantitative, should have a clear and recognizable statement of aim(s).

Selection of an Appropriate Research Strategy

In formulating the aims of the investigation, the researcher must consider the best research strategy. For instance, if the demonstration of causal effects is required (e.g. the treatment caused the improvement), a survey method may be inadequate for the aims of the research. If the purpose of the study is to explore the personal interpretations and meanings of participants, then a qualitative strategy will be best. Some researchers now advocate mixed designs in which multiple studies using different methods are performed to examine different perspectives of the same issues. Essentially, the critical appraisal of research is an examination of the extent to which researchers have selected the appropriate research method.

Selection of Appropriate Variables/Information To Be Collected

In a quantitative study, if the selection of the variables is inappropriate to the aims or questions being investigated, then the investigation will not produce useful results. Similarly, in a qualitative study, the information to be collected must be appropriate to the research aims and questions.

Critical Evaluation of the Methods Section in a Research Paper

A well-documented methods section is a necessary condition for understanding, evaluating and perhaps replicating a research project. In general, the critical evaluation of this section will allow a judgement of the validity of the investigation to be made.

Research Participants

This section shows if the study participants were representative of the intended target group or population and the adequacy of the sampling model used.

Sampling Model Used

In [Chapter 5](#), we outlined a number of sampling models that can be employed to optimize the representativeness of a study sample. If the sampling model is inappropriate, then the sample might be unrepresentative, raising questions concerning the external validity of the research findings. In qualitative research, although the participant sampling method may be less formal than in a quantitative study, the issue of representativeness is still pertinent in terms of being able to apply the results.

Sample Size/Number of Participants

Use of a small sample is not necessarily a serious problem, provided the sample is representative. However, given a highly diverse population, a small sample may not be adequate to ensure representativeness of the study sample compared with the population. A small sample size decreases the power of the statistical analysis in a quantitative study. However, in quantitative sampling procedures, sometimes there is disagreement among qualitative researchers as to the issue of how many participants are needed to ensure validity of the study.

Description of the Study Participants

A clear description of key participant characteristics (e.g. age, sex, type and severity of condition) should be provided. When necessary and possible, demographic information concerning the population from which the participants have been drawn should be provided. If not, the reader cannot adequately judge the representativeness of the sample.

Instruments/Apparatus/Tools

The validity and reliability of observations and/or measurements are important for good research. In this section, the investigator must demonstrate the adequacy of the tools used for the data collection.

Validity and Reliability of Research Measurement Tools

A quantitative investigator should use standardized tools or establish the validity and reliability of new tools used. A lack of proven validity and reliability will raise questions about the adequacy of the research findings. A full description of the structure and use of novel tools should be presented so that they can be replicated.

Procedure

A full description of how the investigation was carried out is required for both replication and for the evaluation of its internal and external validity. The requirement of providing a clear and explicit rationale for the research applies to both qualitative and quantitative studies.

Adequacy of the Research Design

It was stated previously that a good design should minimize alternative conflicting interpretations of the data collected. For quantitative research aimed at studying causal relationships, poor design will result in uncontrolled influences by extraneous variables, muddying the identification of causal effects. In [Section III](#) of this text, we examined a variety of threats to internal validity which must be considered when

critically evaluating an investigation. In a qualitative study the theoretical approach taken in the study design or approach should be clearly stated.

Control Groups in Causal Research

When the aim is to identify causal factors in quantitative research, a common way of controlling for bias and confounding effects is the use of control groups (e.g. placebo, no treatment, conventional treatment). If control groups are not employed, then the internal validity of the investigation might be questioned. Also, if placebo or untreated groups are not present, the size of the effect due to the treatments might be difficult to estimate.

Participant Assignment

When using experimental designs such as RCTs, care must be taken in the assignment of participants so as to avoid significant initial differences between treatment groups. Even when quasi-experimental or natural comparison strategies are used, care must be taken to establish the equivalence of the groups. Whenever possible, intervention studies should use double- or single-blind procedures. If the participants, researchers or observers are aware of the aims and predicted outcomes of the investigation, then the validity of the investigation will be threatened by bias and expectancy effects. In qualitative research, it is very important that the research findings are not unduly influenced by the personal positions of the researchers in a way that obscures the meanings and interpretations of the research participants. Of course, the position of the researcher in any study, whether qualitative or quantitative, will to some extent influence the findings but this needs to be kept to a minimum.

Treatment Parameters

It is important to describe all the treatments or interventions given to the different groups. If the treatments differ in intensity, or the administering personnel take different approaches, the internal validity of the project is threatened. The adherence of the study in the delivery of the intervention to the intended intervention is sometimes called *treatment fidelity*. In studies aiming to demonstrate the efficacy, the sequence of any treatments and

observations must be clearly indicated, so that issues such as confounding variables and bias can be detected. Identification of variability in treatment and observation times can influence the internal validity of experimental, quasi-experimental or $n = 1$ designs, resulting in inaccurate decisions.

Settings

The setting in which a study is carried out has implications for external (ecological) validity. An adequate description of the setting is necessary for evaluating the generalizability of the findings. The context of the investigation may have important effects on the study outcomes. Research conducted in the investigator's laboratory or office may yield different results to the same work conducted in the field. The accurate description of settings is a particularly important consideration when appraising qualitative research. The meaning of personal experiences is very much dependent on time, place and culture.

Critical Evaluation of the Results Section in a Research Paper

The results of quantitative studies are analysed using descriptive and inferential statistics. Inadequacies in statistical analysis result in misinterpretation of the research findings.

Tables and Graphs

Data should be correctly tabulated or drawn and adequately labelled for interpretation. Complete summaries of all the relevant findings should be presented.

Selection of Statistics

Where appropriate, both descriptive and inferential statistics must be selected according to specific rules. The selection of inappropriate statistics could distort the findings and lead to inappropriate inferences.

Calculation of Statistics

Clearly, both descriptive and inferential statistics must be correctly calculated. The use of computers generally ensures this, although some attention must be paid to gross errors when evaluating the data presented. Effect sizes and confidence intervals are essential statistics for evaluating the clinical significance of the results. If these statistics are not provided the reader may be unable to apply the results to a practical setting.

Methods of Qualitative Analysis

The methods chosen must complement the theoretical approach taken in the study and be performed according to the specified protocols.

Critical Evaluation of the Discussion Section in a Research Paper

In the discussion section, the investigators interpret the evidence with reference to the initial aims, research question and hypotheses of the investigation. Incorrect interpretation of the evidence may lead to distortions of knowledge and ineffective practices being offered to clients.

Drawing Correct Inferences From the Collected Information/Data

The inferences from the collected information or data must take into account the limitations of the study and the analytical methods used to analyse them. In the quantitative domain, we have seen, for instance, that correlations do not necessarily imply causation, or that a lack of significance in the statistical analysis could imply a type II error or incorrect missing of a real trend or finding. In the qualitative domain, the findings must follow reasonably from the information collected in the investigation according to the paradigm used.

Logically Correct Interpretations of the Research Findings

Interpretations of the findings must follow from the information collected, without extraneous evidence being introduced. For instance, if the investigation used a single-participant design, the conclusions should not claim that a procedure is generally useful for the entire population.

Research Protocol Deviations

In interpreting the data or information collected in a study, the investigator must indicate, and take into account, unexpected deviations from the intended research protocols. For instance, in a quantitative study a placebo/active treatment code might be broken, or 'contamination' between control and experimental groups might be discovered. In a qualitative study, it could be that participants have conversed with each other about the

research before one of the participants completes participation. Investigators are obliged to report such deviations if they are discovered so that the implications for the results might be taken into account.

Generalization From the Findings To the Population

Strictly speaking, the data obtained from a given sample are generalizable only to the population from which the participants were drawn. This point is sometimes ignored by investigators and the findings are generalized to participants or situations which were not considered in the original sampling plan. Qualitative researchers may vary in their willingness to claim generalizability of their findings outside the actual research participants, but this must also be systematically considered.

Statistical and Clinical Significance of the Results

In quantitative studies, obtaining statistical significance does not necessarily imply that the results of an investigation are clinically applicable or useful. In deciding on clinical significance, factors such as the size of the effect, side effects and cost-effectiveness, as well as value judgements of patients concerning outcome, must be considered.

Theoretical Significance of the Evidence

It is necessary to relate the results of an investigation to previous relevant findings that have been identified in the literature review. Unless the results are logically related to the literature, the theoretical significance of the investigation remains unclear.

Summary

A critical approach to appraisal refers to making *judgements* concerning the validity of research projects and evaluating their contribution to knowledge and practice. Our judgements are based on the application of the principles of research to analysing a specific research paper or as a set of related research papers. This is obviously the reason why you need to understand the basic principles before you can conduct a valid critique. The critical approach is not at all negative or destructive but is an essential component

of constructing knowledge based on the best available evidence. Being capable of conducting a methodological critique is an essential skill for practitioners committed to evidence-based health care.

Even when some problems are identified with a given research report, it is nevertheless likely that the report will provide some useful additional knowledge. Given the problems of generalization, an individual research project is usually insufficient for firmly deciding upon the truth of an hypothesis or the usefulness of a clinical intervention. The reader needs to scrutinize the range of relevant research and summarize the evidence using qualitative and quantitative review methods. In this way, individual research results can be evaluated in the context of the research area. Disagreements or controversies are ultimately useful for generating hypotheses for guiding new research and for advancing theory and practice.

References

1. Hicks C.M. Applied project design and analysis. In: Hicks C.M, ed. *Research Methods for Clinical Therapists: Applied Project Design and Analysis* . 5th ed. London: Elsevier Health Sciences; 2009.
2. Dawes M, Davies P, Gray A, Mant J, Seers K, Snowball R. *Evidence-Based Practice: A Primer for Health Care Professionals* . 2nd ed. Edinburgh: Churchill Livingstone; 2005.
3. Soares-Weiser K, Thomas S, Thomson G, Garner P. Ribavirin for Crimean-Congo hemorrhagic fever: systematic review and meta-analysis. *BMC Infect Dis* . 2010;10:207.

SECTION 7

Evaluation and Dissemination of Research Results

OUTLINE

Introduction

23. Synthesis of Research Evidence: Systematic Reviews and Meta-Analyses

24. Translational Research

Introduction

As discussed in [Chapter 3](#), having completed a research project, the researchers need to communicate their research results to the community of health scientists and health professionals. A common way of communicating research findings by established researchers is first to report the results at a professional conference and to then write a more formal publication in a peer reviewed journal. There are many thousands of peer reviewed journals. Each journal has its particular set of rules and requirements for how research projects should be written up for publication. In general, at least for quantitative research, the format for presenting our research follows the sequential stages of the research process outlined in this book. This general format is outlined in [Chapter 3](#), which includes a detailed discussion of the specific sections of a research paper and outlines some ‘stylistic’ considerations required by journal editors.

In [Chapter 23](#), we outline how we evaluate the quality of a research paper and its results. We also discuss what happens if we find serious problems with the design, data collection and analysis and interpretations of a research project.

We discussed in [Chapter 1](#) how, as clinicians using evidence-based health care, we need to be able to critically evaluate research publications. For clinical researchers it is an ethical requirement that we report our results in an accurate and honest fashion. Before a paper is published in a reputable journal, it is critically evaluated by experts in the area (called *referees* or *peer reviewers*) for errors or problems. However, although reviewers do their best, sometimes serious problems can remain unidentified. Ultimately, it is our task as health professionals to read important publications critically. We owe it to our patients and clients to be cautious and critical concerning the research we use to back up our clinical practice. However, being critical

does not mean we must be cynical or derogatory toward the work of health researchers. It is difficult to do good-quality research.

In reality, a single research project is rarely sufficient either to verify or falsify a theory, or to convincingly demonstrate the effectiveness of a treatment. Rather, we need to evaluate and summarize the literature as a whole: that is, conduct a systematic literature review. Conflicting findings or gaps in the knowledge for a given area of health care identified in our literature review encourage further research. In [Chapter 24](#), we discuss how systematic reviews and meta-analyses are conducted in health sciences. Well-conducted systematic reviews provide the foundations for evidence-based health care.

Synthesis of Research Evidence

Systematic Reviews and Meta-Analyses

Introduction,
Basic Principles,
Narrative Reviews of Research Evidence,
Systematic Reviews,
 Steps in Conducting a Systematic Review,
 Research Questions,
 Theoretical Framework for Systematic
 Reviews,
 Search Strategy for Systematic Reviews,
 Appraisal and Selection of Key Papers for the
 Review,
 Coding the Included Research Studies,
 Interpretation of the Research Evidence,
 Cochrane Systematic Reviews,
Meta-Analysis,
 Extracting Data From Individual Studies,
 Combining Data From Diverse Studies,
 Interpreting the Results of a Meta-Analysis,
Validity of Systematic Reviews,
 Sampling,

Extracting the Data,
The State of the Research Program,
Evidence-Based Practice,
Summary,

Introduction

Systematic reviews are now a common research method used in the design and delivery of health care. This chapter describes the important concepts and methods in systematic reviews and provides practical examples from the published literature. It also describes how to conduct a systematic review and how to use the results in clinical practice. We draw on examples of systematic reviews published in a range of international journals and databases to illustrate the principles.

The research papers published in scientific and professional peer-reviewed journals are the most important source of research for driving evidence-based health care. Each research publication contributes to the overall knowledge base for describing, understanding and solving health problems. However, an individual research project is usually insufficient for firmly deciding the truth of a research hypothesis or determining the efficacy of a clinical intervention. To have more soundly based evidence, we need to compare, contrast and synthesize the results of research papers dealing with the same topic. Consistent results across diverse research studies provide a sounder basis for justifying the evidence-based delivery of health services. As noted by Masic et al: ‘Evidence-based medicine (EBM) is the conscientious, explicit, judicious and reasonable use of modern, best evidence in making decisions about the care of individual patients’. ¹ (p219)

A systematic review is a report that ‘summarises the results of available carefully designed healthcare studies (controlled trials) and provides a high level of evidence on the effectiveness of healthcare interventions’. ² This information can be used by health authorities to allocate resources and to determine treatment standards by clinicians to inform their treatment choices, and by patients and families to assist in their decision making.

Systematic reviews are useful because in most research topics there are multiple studies with different outcomes. So which study should be followed? Single research studies are no longer accepted as sufficient evidence to guide critical clinical decisions. Evidence-based medicine and evidence-based public policy require good quality evidence in the form of systematic reviews.

In the past, reviews of research evidence associated with a particular topic were often unstructured and not rigorous. These narrative reviews were essays that were based on the researcher's personal reflections of the evidence. There were no strict protocols concerning how research studies were selected for inclusion in the review and also which studies should be excluded from the review. There were also no specific requirements for the reporting of the results of the review. Thus the quality and detail of the early reviews varied widely. This was an unsatisfactory situation in that the reviewers could intentionally or unintentionally bias the results of the review. This led to the need for 'systematic' reviews where the decisions made by the reviewers are defended and unbiased scientific standards are available for independent scrutiny.

The term '*research literature*' refers to the set of publications containing the network of theories, hypotheses, practices and evidence representing the current state of knowledge in a specific area of the health sciences. A literature review contains both the critical evaluation of the individual publications and the identification of emergent trends and patterns of evidence based upon these studies. The literature review is a synthesis of the available knowledge in an area and therefore constitutes the strongest foundation for initiating further advances in theory and practice.

The overall aim of this chapter is to outline the basic strategies used for synthesizing evidence and producing a literature review.

The specific aims of the chapter are to:

1. Identify the basic methodological principles relevant to writing a health sciences review.
2. Describe the process for conducting a narrative review.
3. Describe the process for conducting a systematic review.
4. Explain how to interpret the findings of a published meta-analysis.

5. Explain the uses and limitations of systematic reviews and meta-analyses.

Basic Principles

The first thing to consider is that writing a literature review is a demanding intellectual challenge. The facts do not ‘speak’ for themselves. Rather, the evidence has to be extracted, critically evaluated, organized and synthesized into a logical, coherent representation of the current state of knowledge. For example, consider the review by Olanow titled ‘The scientific basis for the current treatment of Parkinson’s disease’. ³ This relatively brief, 15-page review is based on only 75 references, although there are thousands of research papers, articles and reports available on the anatomy, physiology and treatment of Parkinson’s disease. In writing the review, the author had to make a series of expert judgements regarding which were the nine key papers, containing the most salient, up-to-date information for understanding the medical treatment of Parkinson’s disease. Many papers were rejected.

Second, the outcome of the review process is influenced by the theoretical orientation and professional background of the reviewer. Olanow, a leading neurologist, provided an authoritative review written from a biomedical perspective. ³ In contrast, a physiotherapist working in neurologic rehabilitation might take a different conceptual approach to the causes and treatment of Parkinson’s disease. He or she might place more emphasis on psychological and social factors as integral components of the aetiology and treatment of Parkinson’s disease.

Third, the selection, analysis, critique and synthesis of the materials is an active, interpretive process drawing on the personal experiences, interests and values of the reviewer. Even if their professional backgrounds are identical, there are no guarantees that two reviewers interpreting the evidence from the same set of publications will arrive at exactly the same conclusions. Depending on how these reviewers approach the subject matter, they might emphasize different aspects of the evidence or select different patterns in the data as being important, resulting in different syntheses. In [Chapter 1](#), we discussed the postpositivist position that

theories and preconceptions can influence our perceptions of what is happening in the world, and therefore shape the way in which we construct knowledge.

Last, the notion that we all have our experiences, opinions and prejudices does not imply that ‘anything goes’ when writing health sciences reviews. On the contrary, we need to apply the principles of the scientific methodology to ensure that we provide an accurate overview of the literature. In other words, there are principles that we must follow in preparing a literature review.

As stated previously, in preparing literature reviews and evaluating research findings, a multiplicity of papers must be considered, at least according to the general steps listed here:

1. Identify relevant literature; select key papers.
2. Critically evaluate key papers, as discussed in this section. You might decide to discard some papers if irreparable problems are discovered.
3. Identify general patterns of findings in the literature. Tabulate findings, where appropriate.
4. Identify crucial disagreements and controversies.
5. Propose valid explanations for the disagreements. Such explanations provide a theoretical framework for resolving controversies and proposing future research.
6. Provide a clear summary concerning the state of the literature, identifying progress, obstacles and further research.

Narrative Reviews of Research Evidence

There are several approaches to conducting health sciences reviews. For example, the previously mentioned review by Olanow can be classified as a ‘narrative’ review.³ This approach entails producing a ‘story-like’ overview of the state of current evidence and theories. A narrative approach to reviewing literature means that the review is conducted in a storytelling fashion. There are basically two kinds of narrative reviews: those where the review constitutes the entirety of the paper, and those which are only part of the paper, and are integrated into the introductory section of a paper that

reports a research study. Narrative reviews have an important place in the evidence base for health practice. For instance:

- A narrative review is an essential part of the 'Introduction' section of all reports of research studies. The review outlines the current state of knowledge, identifies gaps in that knowledge, justifies the research in terms of its capacity to advance knowledge and practice, and provides that background and the rationale for the research questions.
- A narrative review is also an essential first step for constructing models and theories. The review is used to identify and synthesize available evidence and then to demonstrate the 'fit' between the proposed conceptual framework (theory) and the facts.
- A narrative review also can be useful in identifying current best practice in preventing and treating a particular health problem.

The quality of narrative reviews varies, given that the analysis, critique and synthesis of the material is an active process, which draws on the creativity and intellectual style of the reviewer. There is no guarantee that two people interpreting the evidence from the same set of papers will produce identical syntheses. Depending on how individuals approach the subject matter of a field, they may emphasize different aspects of the evidence or select different patterns in the data, resulting in a potentially different overall conclusion. In the extreme, some health workers hold biased theoretical and ideological positions to such an extent that the meaning of the evidence could be completely distorted (e.g. a medical scientist employed by a tobacco company, who reviews literature relating to the health risks). There are various sources explaining how to conduct reviews (see <http://www.prisma-statement.org/>). Narrative reviews for the purpose of systematic evidence review are now essentially superseded by the use of systematic reviews. Many journals do not accept narrative reviews as acceptable evidence reviews.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA [<http://www.prisma-statement.org/>]) is the global reporting standard for systematic reviews and meta-analyses.⁴ PRISMA is focused on randomized controlled trials (RCTs) and intervention studies. The

PRISMA statement consists of a checklist and flow diagram that is followed to perform and report the review. It is now mandatory for many journals. Systematic reviews have to meet PRISMA standards to be considered for publication in many journals.

PRISMA also recommends the prior registration of the protocol for the systematic review via the PROSPERO (International Prospective Register of Systematic Reviews). ⁵ PROSPERO is an international systematic review registration system that is funded by the United Kingdom National Health Service National Institute for Health Research. However, it includes registrations from all countries, so it is an excellent source to find relevant reviews.

Systematic Reviews

Although reviewers generally adhere to the principles of science and logic in conducting a narrative review, there have been concerns about bias in the selection of evidence to be included and lack of clarity. More recently, systematic reviews and meta-analyses have been introduced to enhance the rigour for combining and interpreting the state of the literature. Systematic reviews rely on the explicit use of the methodological principles discussed in previous chapters. In effect, systematic reviews follow the problem-solving approach as used for conducting empirical research as outlined in [Chapter 3](#). Let us examine a published example ⁶ to illustrate the logic and principles underlying the conduct of systematic reviews.

Steps in Conducting a Systematic Review

There are many references that describe how to conduct a systematic review. The PRISMA documentation is excellent. Khan et al provide a simple explanation of the basic steps: ⁷

Step 1: Framing Questions for a Review

The review research questions must be clear, unambiguous and structured. There are good resources to assist reviewers with developing clear questions. The PICO(T) system is useful. ⁸

- **Population:** refers to the participants in the studies.
- **Intervention:** refers to the treatments.
- **Comparison:** refers to what you will use as a reference group to compare with the intervention.
- **Outcome:** refers to what is measured.
- **Time:** describes the duration of the data collection.

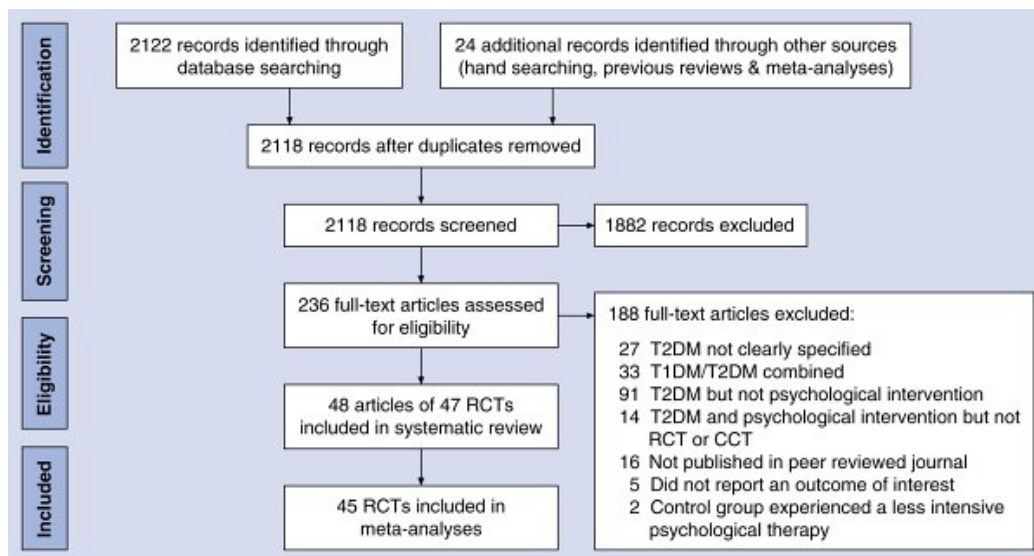


FIG. 23.1 PRISMA flow chart review of diabetes interventions based on randomized controlled trials conducted in China.

Step 2: Identifying Relevant Work

A successful search strategy must include a specification of the sources and databases to be searched and the search terms to be used to find the studies. There needs to be clear statements of inclusion and exclusion criteria for the studies. For example, you may decide to include only RCTs. (See Polgar and Thomas, 2018 ⁹, for a comprehensive description of key study types.)

Step 3: Assessing the Quality of Studies

The collected literature that satisfies the initial inclusion criteria will then be subjected to a quality assessment. There are excellent checklists to assist with this. It is essential to check with the journal you intend to publish in concerning which system they favour and to follow it. The Equator

Network (<http://www.equator-network.org>) is an international collaboration that has useful decision guides to select the right tools to perform study quality assessment.

Here, we present an example of a PRISMA flow chart from one of our team's systematic review publications.¹⁰ This is a review of diabetes interventions based on RCTs conducted in China. Fig. 23.1 outlines how the search progressed and what studies were finally included. You will note that 2122 studies were initially identified but only 45 were included in the analysis. Actually, this is a relatively high included number. It is interesting that many studies are rejected for quality reasons in systematic reviews.

Step 4: Summarizing the Evidence

This involves the tabulation of study characteristics, quality and effects. The tables are typically included in the publication or in a web-based link to the paper. A key consideration is assessment of research 'bias' for all the studies. 'Bias' means good or poor quality. Many studies will be rejected from the analysis. The meta-analysis is then performed on the remaining studies. This is a statistical method for combining the results of multiple studies. Additionally, a narrative descriptive analysis of the studies is provided.

Systematic reviews often include extensive tables describing the studies' results.

Fig. 23.2 is from one of our team's meta-analyses assessing the effectiveness of motivational interviewing (MI) in controlling hypertension in China.¹¹ This figure is referred to as a forest plot, which is a graphical representation of a meta-analysis.

The data on the left shows the individual study results. The diagram on the right is a graph of the meta-analysis results. These data show a small positive effect of the MI interventions on diastolic blood pressure across the reported studies (mean difference -0.90 ; 95% CI -1.75 to -0.06) and no significant overall effect for diastolic blood pressure follow-up studies (mean difference -0.58 ; 95% CI -1.25 to 0.09). There is also the issue of heterogeneity of results. The heterogeneity of study results is measured by various statistics and it provides a measure of the consistency of the various study results.

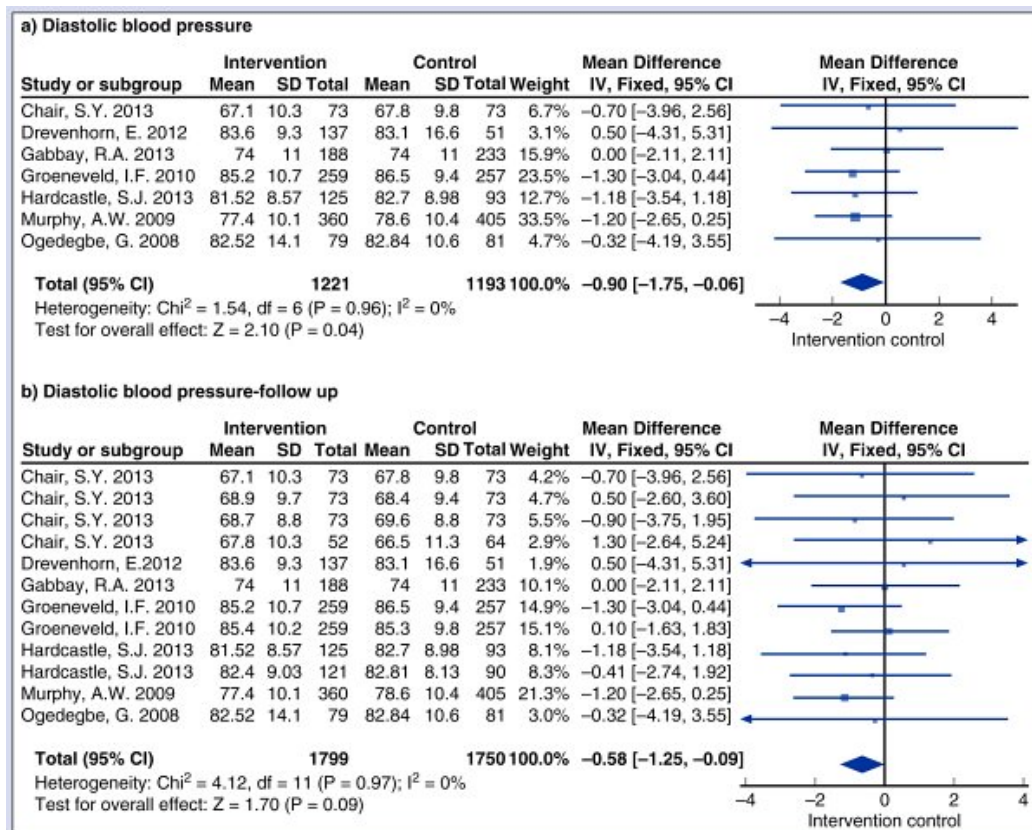


FIG. 23.2 Forest plot meta-analyses assessing the effectiveness of motivational interviewing in controlling hypertension in China.

Step 5: Interpreting the Findings

This involves the formulation of recommendations based on the evidence. The recommendations are graded according to the strength of the evidence. It is important that the findings are a sober and accurate assessment of the current evidence.

Research Questions

When conducting a systematic review, we are expected to formulate a clear review research question that will be answered by the outcomes of the review. You might have read about the research using embryonic or stem cells for reconstructing the brains of people suffering from neurologic conditions such as stroke and Parkinson's disease. We were interested in reviewing the evidence for answering the research question 'How effective

is reconstructive neurosurgery (i.e. the grafting of immature cells) for improving the signs and symptoms of Parkinson's disease?'.⁶

Theoretical Framework for Systematic Reviews

Although this principle is not adhered to by all reviewers, it is very useful to specify the theoretical framework(s) which guides a specific review. The theoretical framework used for conducting the review was identified as the 'Repair Model'. This is a purely biomedical, quantitative view of neural reconstruction based on the notion that recovery is due to the replacement of dopaminergic cells damaged in Parkinson's disease. An explicit theoretical framework is essential for understanding a given area of health as a coherent research program.

Search Strategy for Systematic Reviews

The next step is to identify the relevant publications. Several issues are relevant to devising a search strategy. First, a number of different sets of key terms are used to conduct the electronic search to 'match' these terms with those in the papers in the different online databases (e.g. Medline, Embase, CINAHL etc.). A 'hand search' of current conference proceedings is conducted as a double-check for key authors. In general the reviewer should have available at least two to three key papers that can be included in the review. For example, in Polgar et al⁶ we conducted the search shown here:

A search of Medline (1994–2000 and 2000/01–2000/10) using the exploded terms 'fetal tissue transplantation', 'Parkinson's disease, human or fetal tissue transplants', 'Parkinson's disease, human', also using the author 'Kopyov' was conducted. In conjunction with this search, abstracts from the American Society for Neural Transplantation and Repair conferences (1999, 2000) were hand searched for authors that may have been overlooked. The reference lists of key papers were searched to identify papers that might not have been identified through online search mechanisms. An additional electronic search was conducted using the following databases: Medline

(1966–March Week 2, 2001), Embase (1994–April Week 1, 2001), CINAHL (1982–March Week 2, 2001).

The general point here is that reviewers must be diligent in identifying all the publications which constitute the literature in the area targeted for review. The search should include both ‘electronic, online searches’ and ‘hand’ searches of key journals for cross-checking if the relevant papers were identified by the search engines. It is essential to have a working knowledge of who the key researchers are in a particular field and what critical issues exist in a research program before we can undertake a formal published review.

Appraisal and Selection of Key Papers for the Review

Critical appraisal refers to applying the principles of design, sampling, measurement and analysis to make judgements about the quality of research evidence (see [Ch. 22](#)). Depending on the area of health sciences being reviewed, literature searches might yield any number of publications, from one or two to hundreds. Relying on the outcome of the search, the reviewers might restate the research questions and redefine the scope of the proposed systematic review. This will expand or restrict the number of studies for further searchers. Additionally, explicit inclusion/exclusion criteria are used to identify the most relevant sources.

For example, Polgar et al ⁶ used the following criteria for including a study in a review. To be included in the review, a study had to:

- be published in a peer-reviewed journal.
- have transplantation surgery performed after 1993, after consensus for optimal donor characteristics.
- have grafted human or embryonic cells.
- have followed ‘best practice’ stereotactic surgery procedures.
- have followed standard postsurgical assessment protocols.

It would be tedious in the present context to explain the specific reasons for each of the selection criteria; interested readers will find that they were

justified in the original review. The point is, we must have explicit, objective criteria for including or excluding studies.

The key papers function as a sample of the best available information accessible in the literature. Parallel to empirical research, we use the evidence from key papers to draw inferences about the state of knowledge in the area under review. The use of diagnostic searches and explicit inclusion/exclusion criteria ensure that the ‘sample’ of papers produces a representative, rather than a biased sample of the overall knowledge.

Coding the Included Research Studies

When the process of selecting and collecting the publications is complete, we have the available evidence relevant to answering our research question. The information that we seek is embedded in the text of research publications. We need to extract the key information from the text of each of the papers selected for the review. We can approach the analysis of the meaning of this information in a way that is similar to qualitative data analysis (see [Ch. 22](#)). Similarly to the predetermined codes used in content analysis, reviewers use the constant features of quantitative research for identifying the categories appropriate for analysing the key papers. These features include:

1. The design of the studies.
2. The sampling strategies used and the sample characteristics.
3. The ways in which the treatment or intervention was administered.
4. The data collection strategy used (i.e. the measurement strategy).
5. The statistical and clinical significance of the results.

Other dimensions or features might also be coded or these features can be modified according to the judgements of the authors. These aspects of the research process are often presented in a table form, as shown in [Table 23.1](#). We used a hypothetical example rather than the previous published review.

Table 23.1

Publication	Sample Size	Average Age of Patients (Y)	Method of Measuring Compliance	Patients Compliant (%)
Smith (2000)	50	55	Self-report	85
Jones (2001)	60	58	Self-report	82
Brown (2001)	50	59	Blood sugar level	40
Miller (2002)	55	56	Blood sugar level	35

To illustrate coding, consider a set of four hypothetical studies reporting on levels of diabetic patients' compliance to self-injection of insulin. The results and key features of the hypothetical studies are tabulated in [Table 23.1](#).

[Table 23.1](#) represents how findings from several publications might be tabulated. Key information about each study, as well as the outcomes, is presented in the table, enabling the emergence and demonstration of an overall pattern to be seen.

Interpretation of the Research Evidence

The coding of the hypothetical studies enables the reviewer to answer the research question: 'How compliant are patients with diabetes with using insulin?' The hypothetical summarized evidence illustrates that sometimes no clear overall trends emerge from the tabulated findings.

In the simple hypothetical example shown in [Table 23.1](#), the percentage of compliance reported by Smith ¹² and Jones ¹³ is more than twice that reported by Brown ¹⁴ and Miller. ¹⁵ Clearly, there is an inconsistency in the literature. A possible explanation for this discrepancy might emerge by inspection of [Table 23.1](#). Neither differences in sample size nor the average ages of the patients provide an explanation for the difference. However, the method by which compliance was measured emerges as a plausible explanation. Smith and Jones relied on the patients' self-reports and might have overestimated compliance levels, ^{12,13} in contrast to Brown and Miller, who used a more objective method and found poor levels of compliance. ^{14,15} Of course, this explanation is not necessarily true, but is simply an hypothesis to guide future investigations of the problem. There are other possibilities which might account for the pattern of findings. It appears that more research is needed.

Cochrane Systematic Reviews

Archie Cochrane, Scottish doctor, was an early leader in evidence-based medicine. In 1971 he published his book 'Effectiveness and Efficiency: Random Reflections on Health Services'. The book strongly criticized the lack of rigorous evidence for many medical interventions and the associated waste and poor effectiveness of these. Dr Cochrane's call for improved evidence and clinical effectiveness and efficiency culminated in the establishment of the first Cochrane Centre in 1993. This has grown to an international network of centres that focus on specific topic areas.

The process by which a Cochrane Review is performed contains many of the steps included in other systematic review schemes.¹² However, it has much greater formality and rigour in its processes. The author review team proposes the topic and content to the relevant Cochrane Review Group. If their proposal is accepted, the hosting Cochrane Review Group is heavily involved in providing expert technical advice to the author review team. This ensures high quality in the review. Furthermore, the protocol for the review is publicly released and made available for independent public review before commencing the actual review. Once the comments have been dealt with to the satisfaction of the Cochrane Review Group technical team, the processes continues to the actual review. The Review Group closely monitors the conduct of the review to check its quality.

A Cochrane Systematic Review is a major undertaking. It requires an extended time frame to complete because of the rigorous review and checking processes and the level of dialogue with the technical members of the hosting Cochrane Review Group is often substantial. They are also very lengthy as seen in the Cowlshaw et al¹³ review from our team. However, Cochrane Reviews are recognized as the premier form of systematic review and are highly respected. Dr Cochrane's call for better evidence in health care has been very well met by the international network that carries his name.

Publication of Systematic Reviews

Good systematic reviews generate a lot of citations and are good for the careers of the authors. They are popular with journal editors because they

increase the impact factor of the journals.

Meta-Analysis

Dooley (2001) ¹⁴ discussed the availability of different strategies for summarizing research findings from multiple studies:

1. *Qualitative*. A qualitative review involves the selection of key features of related publications such as designs, participant characteristics or measures used in the studies. These features are presented in a table form, such that differences in the features of the research can be related to outcomes. The qualitative reviews identified by Dooley are related to the systematic reviews, as we discussed previously. ¹⁴
2. *Quantitative*. A quantitative review calls for the condensation of the results from several papers into a single statistic. This statistic represents an overall or pooled effect size. These procedures are meta-analyses which are systematic procedures for summarizing the results published in a set of research papers.

Although many statistical procedures can be used for synthesizing data, meta-analysis also refers to an active area of statistics examining strategies most suitable for synthesizing published evidence. Statisticians have developed software packages such as ‘Comprehensive Meta-Analysis’ which expedite the computational difficulties entailed in synthesizing evidence from diverse studies (<http://www.metaanalysis.com>).

Extracting Data From Individual Studies

The basic requirement for a meta-analysis is extraction of equivalent results from the individual studies. If the studies have different outcome measures, then the meta-analysis is difficult (although not necessarily impossible) to perform. The clearest approach is to extract data which is measured in exactly the same way, across these studies. For example, pain can be assessed in a number of different ways. These include in terms of a dichotomy (e.g. whether the person is experiencing pain or is pain free); by

rating the self-perceived intensity of the pain on a 10-point scale; by rating the intensity on a 100-point visual analogue scale; by measuring the amount of analgesia required to remain pain free. The optimal situation for synthesizing results across two or more related studies on pain is one where the pain outcomes were measured in the same way in each of the studies.

The size of the effect of the intervention (or the effect size) is of primary interest when synthesizing the results of RCTs using a meta-analysis. There are a number of measures of effect size for both dichotomous outcomes (e.g. experiencing pain, pain free) and continuous outcome measures (e.g. ratings of pain on a scale of 1 to 10). Some of these measures of effect size are discussed in the next sections. When synthesizing the results of several RCTs, the same effect size measures are preferable.

Table 23.2

Study	Participants (n)	Average Age (_____)
A	80	40
B	50	45
C	100	60

Combining Data From Diverse Studies

How are results synthesized? Let us look at a simple example for combining data. Say that you are interested in the average age of the participants in three related studies: A, B and C ([Table 23.2](#)).

Say we wish to calculate the average age for all the 230 participants in the three studies. Could we calculate the overall mean, $\bar{x}$, simply by adding up the three means and dividing by three? The answer is no, because there are different numbers of participants across the groups. We must give a weight to each of the statistics depending on 'n', the sample size for each study. The equation which we use is:

$$\begin{aligned}
\bar{X} &= \frac{(\bar{X}_A \times n_A) + (\bar{X}_B \times n_B) + (\bar{X}_C \times n_C)}{n_A + n_B + n_C} \\
&= \frac{(40 \times 80) + (45 \times 50) + (60 \times 100)}{80 + 50 + 100} \\
&= 49.8
\end{aligned}$$

The point here is that, to calculate the correct overall statistics, we must give a weight to each study. In general, the weight assigned to a study represents the proportion of information the study contributes to the overall analysis. For the calculation just presented, weight was determined by the sample size used in each study. In general, the weight assigned to a study represents the proportion of the information contributing to the calculation of the overall statistic.

Even the calculation of a simple statistic like ‘average overall age’ can be useful for understanding the state of a research program. For instance, in Polgar et al ⁶ we found that the mean age of patients with Parkinson’s disease was 56 years, with the mean overall duration of the illness being 13 years. These results indicated that the average age of onset of the disease was only about 43 years, indicating that experimental reconstructive neurosurgery has been offered to an unrepresentative sample of patients with Parkinson’s disease. Typically, people with Parkinson’s disease are in their late 60s or early 70s.

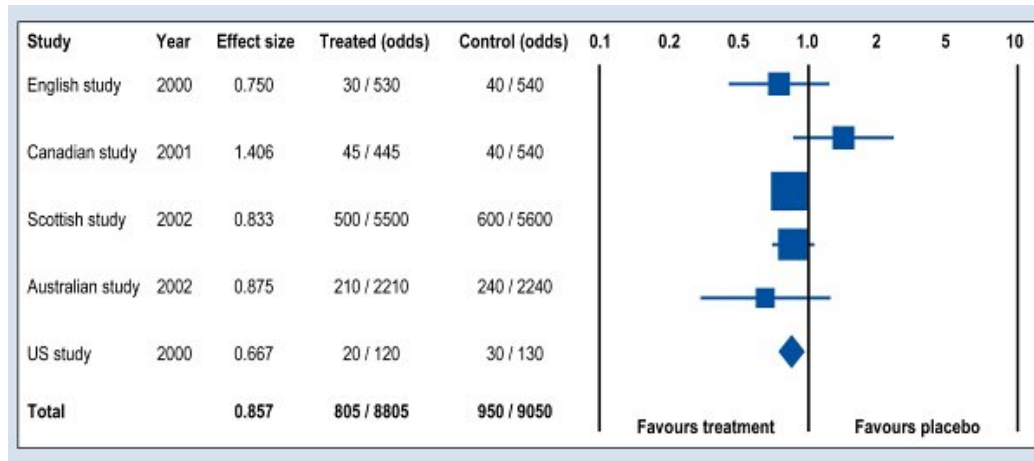


FIG. 23.3 Results of a hypothetical meta-analysis: negative findings.

Of course, we need to synthesize other clinically and theoretically relevant statistics appearing across the papers constituting a research program, including the overall pooled standard deviation, the overall statistical significance, the overall effect size and the confidence intervals (CIs) for the overall effect sizes. These analyses are best carried out using statistical software packages. You can check the Internet for further information regarding the logic of meta-analysis and currently available software packages.

Interpreting the Results of a Meta-Analysis

There are different ways for conducting and reporting the results of meta-analyses which have become quite frequent in the health sciences literature. A typical way of presenting the results is shown in Figs. 23.3 and 23.4, which show the hypothetical outcome of a computer-assisted meta-analysis (modified from www.metaanalysis.com). The graphics shown in these figures are referred to as *forest plots*. A forest plot is a visual representation of the results of a meta-analysis.

Say that the printout showed the results of five randomized double-blind trials (see Ch. 7) aiming to demonstrate the effectiveness of a vaccine for influenza. In each of the studies, volunteers were given either the vaccine (treated) or a placebo (control) injection. The features shown in Figs. 23.3 and 23.4 are important for interpreting the outcomes of a meta-analysis.

Studies

After searching and critical analysis of the relevant literature, as discussed previously, the reviewer selected the five papers shown. If you look carefully at the sample sizes in each of the groups under treated (odds) and control (odds), you will find that the sample sizes are not equal. If they represent randomized trials, you might ask why the groups were unequal in the studies. Sometimes unequal groups represent people dropping out because of harmful side effects to the treatment.

Effect

The effect size was represented as an odds ratio (OR), which is a commonly used statistic for outcomes measured on a nominal scale. For this measure the outcome is diagnosed with influenza after vaccination or placebo over, say, 6 months: ‘yes’ or ‘no’. Looking at the first (English) study, the odds of ‘yes’ to ‘no’ are 30/530 for the treated group and 40/540 in the control group. The statistical software package computed an OR of 0.750, indicating a slight reduction in the odds of contracting influenza. Note that an OR of 1 means equal odds or no difference at all, whereas decreasing $OR < 1.0$ favours treatment. For example, an OR of 0.5 would indicate that the vaccination halved the odds for contracting influenza in the sample.

Confidence Intervals

As we discussed in [Chapter 18](#), a 95% CI contains the true population parameter at $p = 0.95$. The lines produced from the ‘squares’ containing the sample OR represent the 95% CIs. You can see in [Fig. 23.3](#) that all the confidence intervals for the five studies overlap with 1, indicating that we cannot infer that the studies favour the treatment.

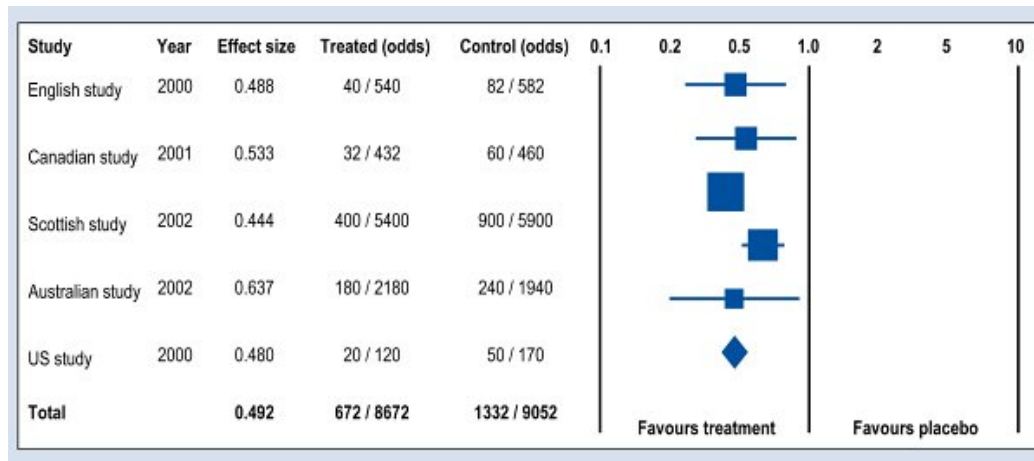


FIG. 23.4 Results of a hypothetical meta-analysis: positive findings.

Weights and Totals

When you look at the ‘odds’ columns in Figs. 23.3 and 23.4, you can see that there is a variation in the sample sizes used in the hypothetical studies. For example, the ‘Scottish’ study included 12 200 participants, whereas the ‘US’ study included only 300 participants. These differences contribute to the relative ‘weight’ of the study represented graphically by the area of the ‘squares’ in the forest plots. Clearly, the larger the square, the greater the sample size. You will also note that the confidence intervals are wider with the smaller squares in comparison to the larger squares. As discussed in Chapter 20, the larger the sample size, the more ‘power’ we have for making accurate inferences.

The ‘Totals’ in Figs. 23.3 and 23.4 refer to the overall statistics synthesized from the results of the five hypothetical studies. These statistics are represented by the diamond shapes on the forest plots. In Fig. 23.3 the total OR was 0.857, indicating a very weak effect for the vaccination. The OR is close to 1 or equal ‘odds’ for having influenza. For the results shown in Fig. 23.4, the total OR was 0.492. This represents a strong effect, indicating that the odds for contracting influenza would have been more than halved by the vaccination. Such results are suggestive of the clinical or practical significance of introducing the treatment.

Validity of Systematic Reviews

The interpretation of the results of published meta-analyses is far more difficult than indicated in the example given earlier. Let us look at some of the sources of difficulties.

Sampling

The results reported in the papers selected for review generally represent a sample of the total information published in a field. This leads us to the first problem: studies that do not report statistically significant findings are often not submitted or accepted for publication. By not having access to these 'negative' findings, the selection of papers becomes biased towards those with reported 'positive' outcomes. Also, in published research papers where outcomes for multiple dependent variables are reported, only the statistically significant outcomes are reported, undermining attempts to synthesize the evidence accurately.⁶

Additionally, the exclusion/inclusion criteria used for selecting the studies can result in a sampling bias. Some practitioners of evidence-based medicine (e.g. Sackett et al ¹⁵) are reluctant to include studies which have not adopted a randomized experimental design. Although this approach has strong methodological justifications, valuable information can be lost by using highly selective inclusion criteria. Of course, the more information that is lost, the weaker the external validity of the review or meta-analysis in relation to the 'population' of research results constituting a research program. The reduced external validity is a trade-off for including only methodologically stronger studies in the review.

Extracting the Data

Another source of error arises when extracting the data from an individual meta-analysis. Some authors report very clear, descriptive statistics but others report their results in an obscure, uninterpretable fashion. Also, some journals and authors only discuss the statistical significance of the results. Obscure and incomplete reporting of the evidence leads to errors in synthesizing overall statistics.

The State of the Research Program

The validity of a systematic review or meta-analysis is limited by the methodological rigour and statistical accuracy of the studies selected for review. To put it bluntly, many health-related problems cannot be resolved and questions cannot be answered on the basis of the currently available evidence. An inconclusive systematic review or meta-analysis is not necessarily a waste of time. Although inconclusive attempts to synthesize data cannot be used to make valid clinical decisions, they provide strong evidence for gaps in knowledge and provide objective grounds for identifying further research required to advance the research program (e.g. Polgar et al ⁶). Until better evidence becomes available we simply provide the best practices suggested by tradition and experience.

Evidence-Based Practice

Although the name of evidence-based medicine or evidence-based practice is relatively new, the idea of using research evidence to inform the design of clinical interventions is very old. Muir Gray provides a very good review of this approach. ¹⁶ Evidence-based health care has been influential in promoting the need for the systematic use of evidence to promote the delivery of high-quality health care (Sackett et al ¹⁵).

As Muir Gray noted, evidence-based practice has three linked ideas at its centre. ¹⁶ These are how to find and appraise evidence, how to develop the capacity for evidence-based decision-making and how to get research evidence implemented in practice. The finding and appraising of evidence draw heavily on systems such as the Cochrane Collaboration approach where evidence is systematically collected and appraised according to predefined principles and protocols. However, the facilitation and implementation elements of evidence-based practice are important additions to the basic establishment of research evidence to support particular approaches in the provision of health services. The evidence-based practice movement is based on the recognition that the mere existence of evidence for the effectiveness of particular interventions does not mean that it will necessarily be effectively implemented. This is the new element of

evidence-based practice: the systematic implementation of programs based on sound research evidence.

Summary

We have seen in this chapter that the advancement of knowledge and practice depends not only on the results of individual research projects but also on the information provided by the synthesis of the results across the literature. Where the projects share the same clinical aims and theoretical frameworks, they are said to constitute a research program. In this chapter, we outlined the process of identifying and selecting publications which contain theoretically or practically relevant research findings. We argued that reviewing a research program is an active, creative process which is influenced by our expectations and attitudes. Although absolute objectivity is not a realistic requirement of a reviewer, there are basic rules to ensure that the review of the evidence is carried out with a minimal degree of bias.

A systematic review of the literature proceeds by identifying the basic components of research papers for organizing the evidence. We use these dimensions to identify patterns or trends which enable us to synthesize the information and answer the research questions. When studies are sufficiently similar, their results can be condensed into single statistics such as overall effect size. In this chapter, we examined how the results of meta-analyses are interpreted.

The relationships between the state of a research program and practice are very complex. We examined hypothetical situations where there was a strong consistency in the research findings and clear trends emerging in the literature could be identified. In these cases, applying the results is relatively straightforward in that we can either adopt or reject the use of a treatment on the basis of the evidence. Systematic reviews and meta-analyses are essential means for identifying best practices available for our patients. However, even well-conceived reviews and meta-analyses can fail to identify clear trends or clinically meaningful effect sizes. When the evidence is inconclusive, we simply continue with traditional practices and identify further research required for resolving unanswered questions concerning improved efficacy. In this way research in the health sciences is

a continuous process, producing better information for advancing theory and practice, as we discussed in [Chapter 1](#) of this book.

References

1. Masic I, Miokovic M, Muhamedagic B. Evidence based medicine - new approaches and challenges. *Acta Inform Med* . 2008;16(4):219–225.
2. Cochrane Consumer Network, . What is a systematic review? <http://consumers.cochrane.org/what-systematic-review>, 2017.
3. Olanow C.W. The scientific basis for the current treatment of Parkinson’s disease. *Annu Rev Med* . 2004;55:41–60.
4. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* . 2015;350:g7647. doi: [10.1136/bmj.g7647](https://doi.org/10.1136/bmj.g7647).
5. Centre for Reviews and Dissemination, University of York, . Prospero: International Prospective Register of Systematic Reviews. 2017. www.crd.york.ac.uk/prospero/.
6. Polgar S, Morris M.E, Reilly S, Bilney B, Sanberg P.R. Reconstructive neurosurgery for Parkinson’s disease: a systematic review and preliminary meta-analysis. *Brain Res Bull* . 2003;60(1-2):1–24.
7. Khan K.S, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. *J R Soc Med* . 2003;96(3):118–121.
8. Riva J.J, Malik K.M, Burnie S.J, Endicott A.R, Busse J.W. What is your research question? An introduction to the

- PICOT format for clinicians. *J Can Chiropr Assoc* . 2012;56(3):167–171.
9. Polgar S, Thomas S.A. *Introduction to Research in the Health Sciences* . 7th ed. London: Elsevier Scientific; 2019.
 10. Chapman A, Liu S, Merkouris S, et al. Psychological interventions for the management of glycemic and psychological outcomes of type 2 diabetes mellitus in China: a systematic review and meta-analyses of randomized controlled trials. *Front Public Health* . 2015;3:252. doi: [10.3389/fpubh.2015.00252](https://doi.org/10.3389/fpubh.2015.00252).
 11. Ren Y, Yang H, Browning C, Thomas S, Liu M. Therapeutic effects of motivational interviewing on blood pressure control: a meta-analysis of randomized controlled trials. *Int J Cardiol* . 2014;172(2):509–511. doi: [10.1016/j.ijcard.2014.01.051](https://doi.org/10.1016/j.ijcard.2014.01.051).
 12. Higgins J.P.T, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0* . Melbourne: The Cochrane Collaboration; 2011.
 13. Cowlshaw S, Merkouris S, Dowling N, Anderson C.J, Jackson A.C, Thomas S.A. Psychological therapies for pathological and problem gambling. *Cochrane Database of Syst Rev* . 2012;11:CD008937. doi: [10.1002/14651858.CD008937.pub2](https://doi.org/10.1002/14651858.CD008937.pub2).
 14. Dooley D. *Social Research Methods* . 4th ed. New Jersey: Prentice Hall; 2001.
 15. Sackett D.L, Strauss S.E, Richardson W.S, Rosenberg W, Haynes R.B. *Evidence-Based Medicine: How to Practice*

and Teach EBM . 2nd ed. Edinburgh: Churchill
Livingstone; 2000.

16. Muir Gray J.A. *Evidence-Based Healthcare and Public Health: How to Make Decisions About Health Services and Public Health* . 3rd ed. London: Churchill
Livingstone; 2009.

Translational Research

Introduction,
(1) Basic Research,
(2) Clinical Research,
(3) Translational Research,
Why Translational Research?,
 Stages of Translational Research,
Phases of Clinical Research,
 Preclinical Research,
'Lost in Translation': Example of Obstacles to Successful
Translational Research,
 Background,
Summary,

Introduction

The ultimate purpose of applied health research is to discover new ways to improve the health of individuals and the community. In previous chapters, we have discussed how the scientific method is applied to produce the evidence required to identify and solve problems of illness and disability. In this final chapter, we will discuss the three major domains of research,

which are basic, clinical and translational research. Although the focus of the present chapter is translational research, we will see that these domains are closely linked when we conduct applied research.

(1). Basic Research

The purpose of what we refer to as basic or ‘pure’ research is to discover and understand the laws of nature. Health or medical scientists are particularly interested in discovering the workings of the biological, psychological and social processes which constitute the foundations for understanding health, pathology and abnormal functioning. For example, advances in the sciences of microbiology and immunology provided the foundation knowledge for understanding and controlling infectious diseases such as pneumonia, tuberculosis or polio. Previously, the outbreak of epidemics has had devastating effects, not only on the well-being or on survival of individuals and their families, but on whole communities.

(2). Clinical Research

The aim of clinical research is to understand the causes and mechanisms of disease and to devise and evaluate new diagnostic tools and therapeutic interventions to prevent and treat diseases. Also, epidemiologic and public health research is essential for understanding health and illness in the population. ¹ A viable, competitive practice is one which is not only effective and safe, but also accessible and affordable for patients and has a positive effect on the overall health of a given population.

(3). Translational Research

Scientific breakthroughs and technological advances have been used for improving health care for more than 100 years. Consider, for example, the serendipitous discovery of x-rays by Wilhelm Rontgen in the late 1890s. This discovery was the source of the idea for the development of modern diagnostic radiology of the 20th century. Accurate imaging has contributed to saving the lives of innumerable people in every country. Advances in physics and computer science have been translated into more detailed and

accurate methods for clinical imaging such as computed tomography (CT) scans or magnetic resonance imaging (MRI). Translational research is a relatively recent approach to ensuring that technological advances and the discoveries in the basic sciences are expeditiously transformed into best practice health care.

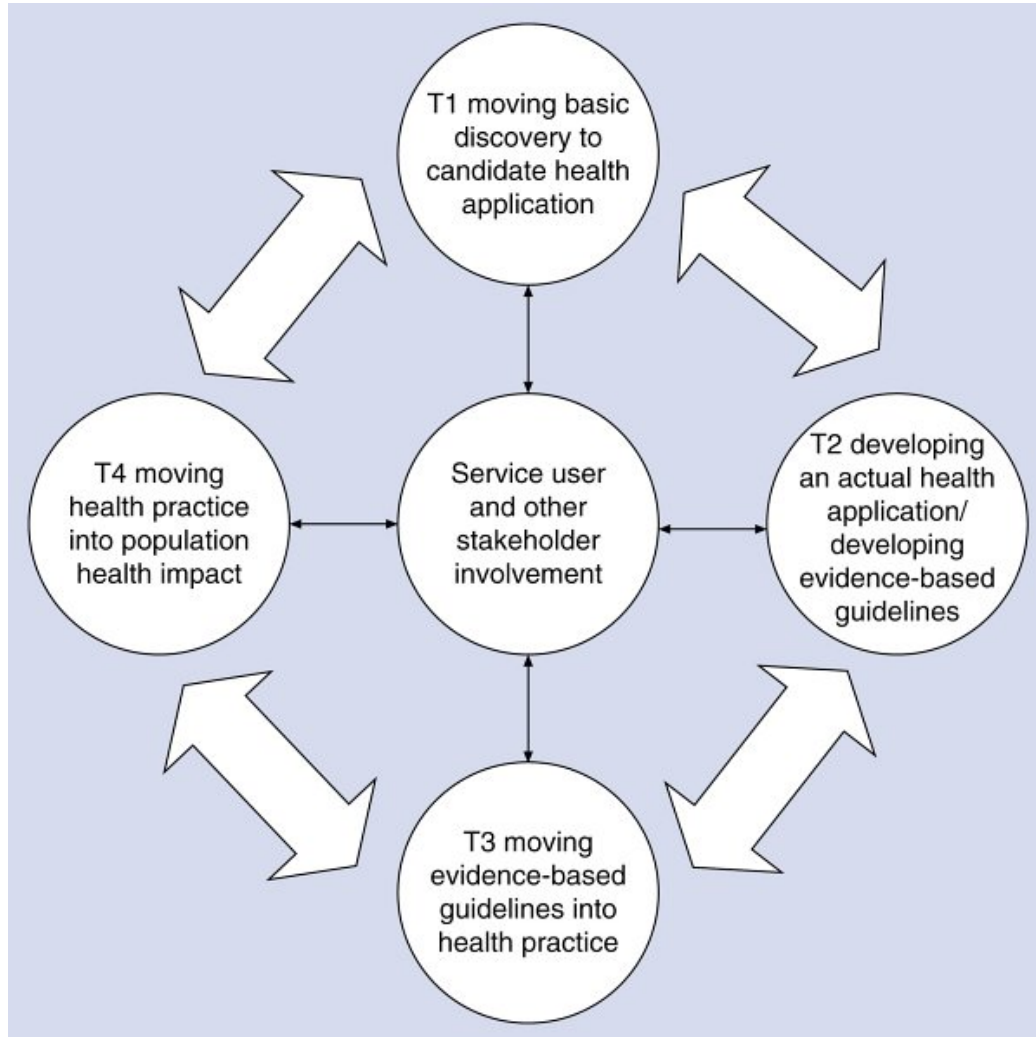


FIG. 24.1 Reconceptualized model of translational research that embeds service user and stakeholder involvement in all phases from Callard F, Rose D, Wykes T. Close to the bench as well as at the bedside: involving service users in all phases of translational research. *Health Expect.* 2011;15(4):389-400.

Translational research has been described simply as the ‘rebranding’ of previous programs of research and development. ² However, it is a very

useful concept to provide you with insight into the rationale of the conceptual framework which guides contemporary medical and health care research. The objectives of this chapter are to:

- Explain the purpose of translational research.
- Describe the phases in conducting translational research.
- Discuss the relationship between the stages of clinical and translational research.
- Identify the obstacles encountered in implementing innovative applied research.
- Outline the principles of a person-centred approach to translating scientific discoveries into effective practices.

Why Translational Research?

The concept of translational research emerged in the late 1990s and subsequently became recognised as an important notion for conducting applied health research.^{2,3} Given the fact that the advances of basic science and technology had been successfully applied well before the 1990s for improving health care, the question is: Why has translational research become so influential at the present time?

One of the reasons for translational research as a key research category was the concern that many promising scientific and technological advances failed to deliver their anticipated benefits.⁴ Using rather colourful language, several commentators argued that a ‘valley of death’ had opened up between basic and clinical research (see [Fig. 24.1](#)). Translational research hopes to cross the ‘valley of death’ by facilitating the process and reducing the frequency of findings being ‘lost in translation’.² The skeletons at the bottom of the chasm represent the failed projects, where apparently useful, even groundbreaking ideas were not actualized as effective health practices. Han et al argued:

The push for increased translational research also stems from ever-present political pressure for the National Institutes of Health (NIH) to

demonstrate that billions of tax-payer dollars are not wasted annually and that NIH-funded research produces tangible societal benefits. ⁵

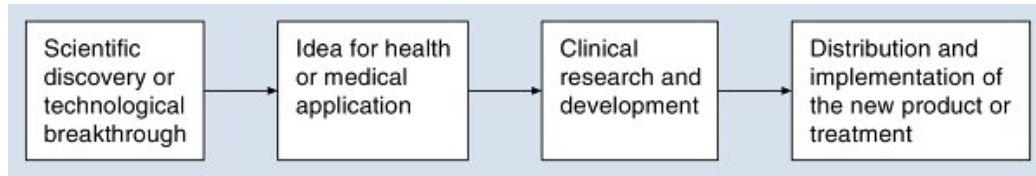


FIG. 24.2 A Typical Research Translation Sequence.

There have been many proposals for improving the outcomes for the translation process. A component of crossing the valley of death has been the injection of billions of dollars into projects aiming to bridge the gap between basic and clinical research. ^{2,6} Apart from increased targeted funding, several specialist professional journals such as *Translational Medicine* were established to publish research which facilitated the integration of basic and clinical research. Although the concept of translational research has become an important component of current health research methodology, there have been controversies in precisely defining the meaning of this term. The translation process is often described as a ‘pipeline’ (i.e. discovery goes in and, if it all goes well, the new treatment comes out; see Fig. 24.1). In the next section, we will look at detailed interpretations of the stages of translational process.

Stages of Translational Research

Several commentators have provided different interpretations of how translational research should be conducted. The simplest version is that it involves moving research findings from the ‘bench to the bedside’. The ‘bench’, of course, refers to the laboratory bench in both a literary and figurative sense. The ‘bedside’ refers to the new treatments which have been made available to be delivered by clinicians to their patients. Therefore translational reach has been characterized as the ‘effective translation of the new knowledge, mechanisms and techniques generated by advances in basic science research into new approaches for prevention,

diagnosis and treatment of disease'.⁶ This simple view is in [Fig. 24.1](#) by the translational pipeline.

There are alternative views of the translation process, particularly relevant to researchers who are seeking to ensure that medical advances are expeditiously implemented. Woolf distinguished between the T1 and T2 phases of the translational research process.⁶ T1 represents the previously discussed 'bench-to-bedside' phase. T2 represents innovative efforts to ensure that a new treatment reaches the patients who are likely to benefit by it. T1 and T2 require different approaches to conducting research. In contrast to the experimental, reductionist methods most suited to T1 research, T2 researchers rely more on descriptive, quasi-experimental and qualitative methods applicable to epidemiology, public policy and qualitative methods. Woolf suggested that insufficient funding was being provided for T2 research so that some of the scientific and clinical advances achieved in the T1 phase were not adequately translated into accessible practices for reducing the burden of disease in the community.⁶ There are numerous reasons why apparently effective practices are not used by sections of the population. An obvious reason is economic; there are expensive products and services which are not affordable to individuals or even poorer countries. The research literature is replete with examples where patients avoid undertaking healthy practices to prevent and better manage especially chronic illnesses.⁷

Khoury et al suggested a more detailed four-stage translation continuum, beginning from basic discovery to implementing improved health outcomes for the community.⁸ [Fig. 24.2](#) represents the suggested four phases of the translation 'pipeline' (see [Fig. 24.1](#)).

To illustrate how the process works, let us look at an example. It has been known for more than 100 years that Huntington's disease (HD) was a genetically transmitted disorder, caused by an autosomal, dominant gene. The onset of HD is around 40 years of age, which means that genetic counselling for people at risk for HD considering having children was problematic. There was a positive attitude by people at risk for HD to predictive tests provided they were safe and accurate. The problem was that in the early 1980s, there were no guarantees that individuals at risk for HD could be given accurate information regarding their genetic status.⁹

T1 Research: From Discovery To Candidate Health Application

The T1 phase involves transforming a scientific discovery into a 'candidate', health application, that is to say, a potential health product or practice. In the early 1990s, the precise location and sequencing of the gene (*Huntingtin*) was identified. The discovery of the gene was applied to developing a 'candidate' biomarker, that is a genetic test for detecting the gene in individuals at risk for HD before the signs and symptoms of HD became apparent. The health application is improved, evidence-based information for enabling individuals with HD to make decisions about their lives.

T2 Research: From Health Applications To Evidence-Based Guidelines

The aim of the T2 phase is to develop an evidence-based guideline to provide for the optimal implementation of the novel practices. In our hypothetical example, the aim is to ensure obtaining accurate data for providing the best possible information to individuals at risk for a genetic disorder, such as HD. That is, the proposed genetic test must be evaluated to demonstrate that it has the characteristics of a good assessment tool. As discussed previously, this implies evidence for properties such as validity and reliability.

T3 Research: From Evidence-Based Guidelines To Health Practices

T3 phase of translational research aims to move the research to the stage where the novel practice is made available for patients. As Khoury et al pointed out, T3 research is a very challenging phase of translational research.⁸ The problem here is that even if the genetic tests are accurate, the findings might not be acceptable to the clients. For example, because there are no effective treatments available for HD at the present time, people at risk for the disease may be reluctant to undergo an accurate genetic test. However, the genetic marker may be very useful to inform them regarding the presence of the *Huntingtin* gene if they were intending

to have children. There has to be both qualitative and quantitative evidence for the benefits of the genetic test for improved quality of life.

T4 Research: From Practice To Population Health Impact

T4, the final phase of the continuum aims to evaluate the ‘real-world’ effects of the interventions. The questions investigated in this phase address the potential social benefits for the translation of the scientific advances. In relation to our example, T4 researchers would seek evidence for the extent to which the genetic tests changed the health and everyday lives of people with HD. Also, public health researchers look for evidence at the extent to which the burden of disease in the population is reduced because of improved genetic counselling. The point here is that applied researchers, who often are also practitioners, need to keep in touch with ongoing scientific advances.

Phases of Clinical Research

Long before the concept of translational research was proposed in the 1990s, the research community referred to the ‘phases of clinical research’ for structuring and organising the phases or steps for developing new health and medical products and practices. The phases for conducting clinical research remain an integral aspect of organizing clinical research, as described on various government and institutional websites (see <https://www.australianclinicaltrials.gov.au/what-clinical-trial/phases-clinical-trials>). Organizing a complex research program into ‘phases’ remains essential for conducting applied health research. As shown in Fig. 24.2, the phases of translational research articulate with the clinical research program.

Preclinical Research

The aim of preclinical research is to test an idea of a clinical development (application) in a laboratory setting. In the context of biomedical research, preclinical research may include the use of animal models of diseases. Animal models enable the preliminary testing the safety and efficacy of a

proposed intervention. As discussed in [Chapter 6](#), there are strict ethical guidelines for the treatment of experimental animals.

Preclinical research provides crucial evidence if the intervention is safe for human research. As an illustration of the steps followed when conducting clinical research, imagine that a laboratory is aiming to develop a new drug (e.g. drug X) for the treatment of an intractable, incurable form of cancer. There is already a massive body of preclinical and clinical knowledge regarding the nature and treatment of cancer which makes the idea of an improved treatment a realistic proposition. The researchers would usually follow the government specified clinical trial phases (see <https://www.australianclinicaltrials.gov.au/what-clinical-trial/phases-clinical-trials>). There are also websites for informing patients who may benefit from drug X about volunteering for clinical trials.

Phase 0 Clinical Trials: This phase or stage involves exploratory or pilot studies with a small sample of healthy human volunteers. The aim would be to test physiological responses to the new drug X. At this stage, small doses of drug X are given to determine whether the drug appears to produce adverse outcomes in the volunteers. Depending on the negative and positive responses there will be a ‘no-go’ decision regarding further development of drug X.

Phase I Clinical Trials: If the decision is ‘go’, the proposed biomedical intervention, such as drug X, is initially tested with a relatively small group of volunteers who have been diagnosed with the target condition. The aim of this stage of research to evaluate the safe dose levels and side effects. In the present hypothetical case, the volunteers would have been diagnosed with the targeted cancer.

Phase II Clinical Trials: Studies are conducted for evaluating the efficacy of biomedical interventions such as drug X. These trials usually employ a variety of research designs, including randomized controlled trials (RCTs) for evaluating the efficacy and safety of the target intervention (drug X). Phase II trials rely on a large number (usually more than 100 volunteers) with the targeted hypothetical cancer being treated. The aim of phase I and phase II trials is to collect sufficient evidence to construct and evaluate an

evidence-based treatment protocol, in this case the hypothetical treatment X for the people with the target disease.

Phase III Clinical Trials: If phase II trials produce clinically meaningful results and the newly developed drug X is shown to be safe for patients, then further RCTs are undertaken. These RCTs aim to compare the safety and efficacy of drug X with currently available treatments for the specific cancer. If these trials are successful in that a clinically significant outcome favours group X, then the new drug may be prescribed for patients.

Phase IV Clinical Trials: Even after a new intervention has been introduced for clinical practice, research continues to optimize its safe and efficient use for treating patients. For example, a small segment of the population with the cancer may respond negatively to drug X. These adverse side effects may occur only in a small proportion of the population, so that the evidence for the problem may become evident only when a large number of patients receive the treatment. Once such subgroups are identified, treatment plans may be modified to exclude them from using drug X.

On a more positive note, new clinical uses for drug X may emerge. New trials may demonstrate that drug X is not only effective and safe for treating the targeted cancer, but also may be effective for other types of cancer. For example, there are drugs, such as aspirin, which have been found over time to have multiple uses ranging from treating headaches to reducing the incidence of strokes.

‘Lost in Translation’: Example of Obstacles to Successful Translational Research

Background

To illustrate the concept of translational research, let us look at a real-life example of the challenges encountered when translating new ideas into effective treatments.

Neural reconstruction is a branch of cellular therapies based on the idea of transplanting cells for replacing neurones lost through degenerative brain

disorders or stroke. Parkinson's disease (PD) is a progressive brain disorder characterized by the loss of dopamine (DA) neurotransmitter expressing neurones located in the basal ganglia system of the forebrain. The anatomic patterns of neuronal loss associated with the motor symptoms and activity limitations are reasonably well understood. PD became the primary target for the development of cell replacement therapies.

A seminal article written by two leading Swedish researchers, Björklund and Lindvall, describes how the different discoveries and theoretical insights into the idea of using DA expressing cells for the reconstruction of impaired brains in PD. ¹⁰

What were the obstacles to translating cell transplantation into a viable treatment for PD?

After more than 3 decades, the idea of using DA-expressing cells was not translated to an effective treatment for PD. Clearly, we can see this represented by the phases of the translational pipeline (Fig. 24.3) where:

- The T1 phase was completed as the discovery of cells was, in fact, successfully translated into an potential health application.
- The T2 phase began successfully with open-label studies indicating that the cells survived expressed DA and helped patients. However, the double-blind studies failed to deliver an evidence-based clinical application of cell therapy. ¹¹
- Therefore T3 and T4 phases were not completed and in this way the project of DA cell transplantation became one of the casualties at the bottom of the 'translational chasm' (Fig. 24.1).

After the failure of the RCTs, researchers had to decide if the work to replace DA expressing cells should continue. Some research groups gave up on the project, whereas other groups (see Barker et al ¹²) decided to proceed with the transplantation of DA-expressing cells.

- We must keep in mind that the successful outcomes of translational research are not only a matter for the scientists and private companies committed to developing and marketing health applications, but for the patients and their families. PD is one of

the most prevalent neurodegenerative conditions; there were great expectations regarding the potential benefits of cell transplantation. It was a great disappointment that the project failed. The question we need to ask is whether the idea is good and realistic. And if it is, what went wrong? Let us look at some of the obstacles: ¹³

- Difficulty of the task: Repairing the human brain using cells emerged as a greater challenge than initially anticipated. There are many factors which interact to influence the outcomes of cell transplantation. These factors include the age of the patient; the severity and duration of the PD; the method of cell preparation; and the optimal source of the DA-expressing cells. In general, great effort is required to complete the translation process from the idea to the point in time when we can deliver a safe and effective treatment for the patients. ¹²

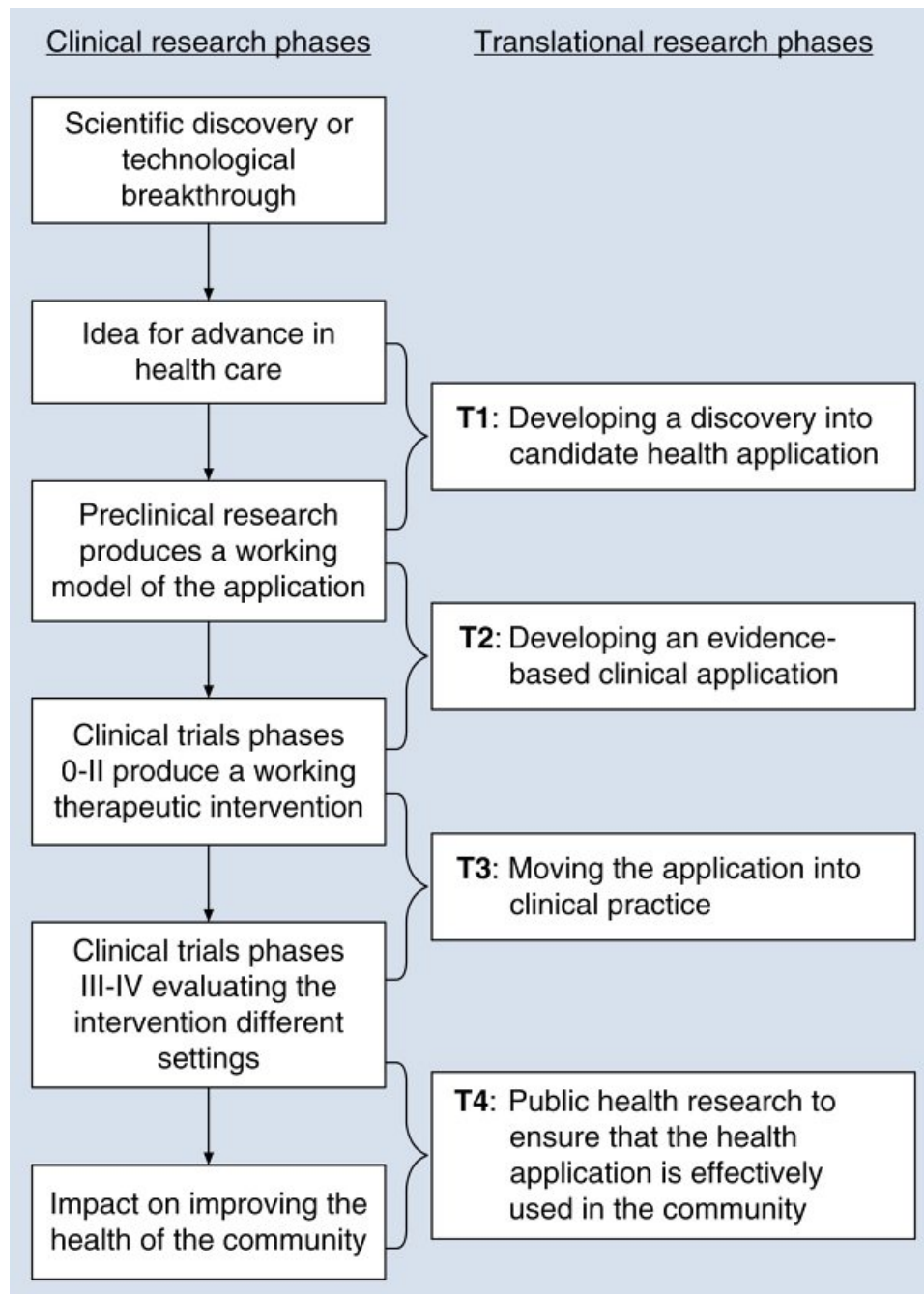


FIG. 24.3 Clinical Research and Translational Phases.

- Changing context for the research project: Another general challenge to developing novel practices is that the overall context for the research project may change over the course of the project. For example, it has become evident that the core neuropathology PD includes progressive impairments in neural systems outside the basal ganglia. Therefore as the understanding of the burden of

disease associated with PD changed, the aims of DA cell replacement therapy shifted from providing an overall cure to the long-term improvement in motor symptoms and improving quality of life.

- Inadequate evidence for the mechanism of action: The preclinical research relied on animal models of PD, generated by neurotoxin-induced lesions. However, these models replicate only a limited aspect of actual PD. In this way, the efficacy of cell transplantation was apparently overestimated, and possible serious adverse outcomes were not predicted by the data. Although considerable caution is required when transferring evidence preclinical research remains an important component of translational research for creating advances in biomedicine.
- Methodological problems: Critical evaluation of the trials revealed many methodological problems may have compromised the translation process. Specific problems have been identified with the accuracy of assessing the efficacy of cell transplantation and the suitability of double-blind randomized trials for confirming the clinical benefits of cell transplantation. It has proposed that, ‘One of the important causes for the diminished efficacy is the reduced opportunity for the participants in double-blind RCTs to learn how to make use of the newly grafted DA-rich cells for achieving optimal recovery’.^{13(p849)}
- Absence of patient-centred approach to translational research: Neural plasticity refers to the overall general mechanism by which the experiences and actions of people ensure the growth and functional integration of new or developing cells into neural systems. It has been suggested that one of the obstacle to achieving optimal recovery has been the inadequate effort by researchers to ensure the involvement of people with PD in the recovery process.¹³ Providing optimal environmental stimulation in the context of a targeted neuro rehabilitation program might enable the level of efficacy required for the successful transplantation of DA-expressing cells as a viable treatment for PD.

In the next and final section, we will describe how the patient centred approach is relevant to implementing translational research.

One of the central themes of this textbook is the importance of adopting a patient-centred approach to conducting health research. In relation to translational research, a patient-centred approach calls for including the active contribution of the participants at all stages of the research process. This approach is represented by Callard et al^{3(p393)} (Fig. 24.1) as shown in Fig. 24.3.

As you can see, Callard et al ³ revised and extended the translational model previously proposed by Khoury et al. ⁸ The reconceptualized model has been modified in two important ways:

1. Rather than a straight pipeline, an ‘interlocking loop’ is proposed.
2. The ‘service users’ are involved in every phase of the translation process.

The ‘interlocking loop’ indicates a continuing process of research and development, as a health application is created, introduced into treatment, then modified through evidence. The ‘service users’ refers primarily to the patients, but there are also other stakeholders, such as patient organizations, health professionals and government regulators involved in the research project as well as biotechnology companies that may be funding the translation process. There are various perspectives on what constitutes appropriate methods and outcome measures to produce a successful treatment. In relation to person-centred research, Fig. 24.1 represents ways in which the research participants may contribute to; and eventually benefits from improving the outcomes of the translational program.

Summary

The goal of translating research findings into proven and widely applied clinical interventions is laudable and necessary. The conversion of investments in expensive clinical research to benefit the general community is now a prime focus in the health and medical research communities. Agencies such as the National Institutes of Health now allocate significant

grant resources to promote research translation. This chapter presented key concepts in translational research, noting that notwithstanding the current emphasis, it has always been the case that researchers and clinicians have been strongly focussed on creating efficient and effective treatments. Now we have formal terminology and research protocols that guide us to the best ways to achieve this. Translational research is an important part of the clinical research landscape and public policy. We could not write our text without canvassing these important issues.

References

1. Rubio D.M, Schoenbaum E.E, Lee L.S, et al. Defining translational research: implications for training. *Acad Med* . 2010;85(3):470–475.
2. Butler D. Translational research: crossing the valley of death. *Nature* . 2008;453(7197):840–842.
3. Callard F, Rose D, Wykes T. Close to the bench as well as at the bedside: involving service users in all phases of translational research. *Health Expect* . 2011;15(4):389–400.
4. Buttery PC, Barker RA. Treating Parkinson’s disease in the 21st century: can stem cell transplantation compete? *J Comp Neurol*. 2014;522:2802–2816.
5. Han X, Williams S.R, Zuckerman B.L. A snapshot of translational research funded by the National Institutes of Health (NIH): a case study using behavioral and social science research awards and Clinical and Translational Science Awards funded publications. *PloS One* . 2018;13(5):e0196545. doi: [10.1371/journal.pone.0196545](https://doi.org/10.1371/journal.pone.0196545).
6. Woolf S.H. The meaning of translational research and why it matters. *JAMA* . 2008;299(2):211–213.

7. Beaglehole R, Epping-Jordan J, Patel V, et al. Improving the prevention and management of chronic disease in low-income and middle-income countries: a priority for primary health care. *Lancet* . 2008;372(9642):940–949. doi: [10.1016/s0140-6736\(08\)61404-x](https://doi.org/10.1016/s0140-6736(08)61404-x).
8. Khoury M.J, Gwinn M, Yoon P.W, Dowling N, Moore C.A, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med* . 2007;9(10):665–674.
9. Teltscher B, Polgar S. Objective knowledge about Huntington’s disease and attitudes towards predictive tests of persons at risk. *J Med Genet* . 1981;18(1):31–39.
10. Björklund A, Lindvall O. Replacing dopamine neurons in Parkinson’s disease: how did it happen? *J Parkinsons Dis* . 2017;7(s1):S21–S31.
11. Olanow C.W, Goetz C.G, Kordower J.H, et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson’s disease. *Ann Neurol* . 2003;54(3):403–414.
12. Barker R.A, Barrett J, Mason S.L, Björklund A. Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson’s disease. *Lancet Neurol* . 2013;12(1):84–91.
13. Polgar S, Karimi L, Buultjens M, Morris M.E. A critical evaluation of the methodological obstacles to translating cell-based research into an effective treatment for people with Parkinson’s disease. *Neurorehab Neural Repair* . 2016;30(9):845–853.

Glossary of research terms

A

AB design A type of experimental design in which the participant is monitored during a baseline phase followed by an intervention phase.

ABAB design A type of experimental design in which the participant is monitored during a baseline phase followed by an intervention phase which, in turn, is followed by further baseline and intervention phases.

Abstract An abbreviated summary of a research report, generally found at the beginning of the report.

Acquiescent response mode A style of answering questions which results in the respondent choosing the middle category in a response scale.

Alternative hypothesis Sometimes, this is also known as the experimental hypothesis. This is the hypothesis for which the researcher is trying to gain support in a statistical analysis, by rejecting the null hypothesis. The alternative hypothesis is represented by the symbol H_A or H_1 .

Apparatus Any equipment or special facilities used in a research project.

Area sample A type of sampling procedure in which the units of the sample are where people live or work, rather than who they are. The researcher divides the target area into sections and then samples the sections.

Assignment The process in an experiment where the researcher allocates subjects to the various groups. Matching and random assignment are the two most common methods. The goal of assignment is to achieve identical groups.

Assignment errors A situation that arises in an experiment where the assignment or allocation of people to groups results in groups with different characteristics.

Authority An appeal to authority argument is based on the proposition that someone of high status knows best, not whether the argument is soundly based.

B

Bar graph A method of displaying data where the frequency of a particular category is reflected in the height of the bar in the graph.

Baseline A phase in an intervention study where the participant is receiving no intervention.

Bell-shaped curve This is the characteristic shape of the normal distribution.

Bias As all humans, health and medical researchers have hopes and expectations regarding the outcomes of a research project. We have an emotional investment in whether the data support our hypotheses or demonstrate that our interventions are effective for our patient. Bias represents the conscious or unconscious ways in which researchers influence participations and the research process resulting in the distortions of the results and drawing erroneous conclusions about the implications of the research project.

Biased sample A biased sample is one that is not representative. It does not reflect the composition of the population to which the researcher is attempting to generalize.

Burden of Disease Burden of Disease (BoD) is a population measure of the effects of a specific disease or health problem

C

Case-control designs These designs involve identifying people with a targeted disease or condition (cases) and comparing them to people who do not have the disease or condition (controls). The aim is to identify relevant health-related differences between the groups.

Causal explanation An attempt to explain the occurrence of a particular phenomenon or event by identifying the cause(s).

Causality An event or factor (A) is generally argued to have caused another one (B) if the following conditions are met: (i) if (A) occurs then (B) occurs; (ii) if (A) does not occur then (B) does not occur; (iii) if (A) precedes (B) in time.

Central tendency The central tendency of a frequency distribution is the average, middle or most common score. Measures of central tendency include the mean, the median and the mode.

Chi-square (χ^2) A statistical test often used with categorical data. It is based on a comparison of the frequencies observed and the frequencies expected in the various categories.

Clinical significance Refers to the practical or applied relevance of the results. Whether or not an outcome is clinically significant is judged by health practitioners, but statistically it is expressed by the effect size. The greater the effect size, the more likely it is that a health-related factor has an impact on clinical decision making such as the selection of an intervention.

Closed-response format A method of eliciting answers from people in a questionnaire in which the researcher provides fixed response categories, e.g. yes or no.

COCHRANE review A form of systematic review of research evidence that is registered with and approved by the Cochrane collaboration. Considered to be a highly rigorous review of research evidence.

Coding A qualitative method of analysis of materials such as interviews where categories are formed and their interrelationships examined.

Cohort studies Two or more groups (cohorts) of participants are compared over a period of time on a number of health-related outcomes. The aim of these studies is to identify factors (exposure) which are different across the groups and may be causally associated with the differences in health outcome. For example, two cohorts may differ on diet cholesterol intake (exposure) resulting in differences in the prevalence of heart disease (outcome). The internal validity of cohort studies is often problematic, as factors other than the identified exposure may be the causes of the outcome.

Complete observer A type of research strategy in which the researcher observes social interactions with no direct personal input; for example, observation via a one-way mirror or through the analysis of a video tape.

Complete participant A research strategy in which the researcher completely participates in the research setting in order to experience its characteristics.

Confidence interval The confidence interval of a sample statistic is the expected range in which the actual population value will be found, at a given level of confidence or probability.

Content validity The extent to which a test or assessment matches the real requirements of the situation' for example, a living skills assessment would have high content validity if it measured cooking, self-care, etc.

Contingency table A method of presenting the relationship between two categorical variables in the form of a table.

Continuous data Data with values that do not fall into discrete categories, e.g. measures of temperature and mass.

Control In an experiment, the researcher attempts to control or eliminate the influence of extraneous variables so that any changes or differences may be attributed solely to the intervention.

Control group In an experiment, a control group is generally a non-treatment group which is compared with the experimental group to study

the effects of the intervention.

Correlation coefficient A statistic designed to measure the size and direction of the association between two variables. The values vary between 0.0 and +1.0 or -1.0.

Correlational studies Studies that are concerned with investigating the associations between variables.

Critical theory In qualitative research, critical theory explains how personal meanings and actions are influenced by the person's social environment.

Critical value of a statistic The value of the statistic (obtained from appropriate tables) that the calculated value for a given result must exceed in order to attain statistical significance.

Curvilinear correlation coefficient A measure of association between variables designed to investigate curved rather than straight-line relationships.

D

DALY A DALY is a year of healthy life that is "lost" because of a specific condition. When the DALYs associated with a condition within a population are summed this is the Burden of that Disease (BoD).

Data The information collected by a researcher.

Declaration of Helsinki The Declaration of Helsinki is an internationally accepted statement made by the World Medical Association concerning ethical principles for human experimentation in medical research. It is regularly reviewed and was originally made in 1964.

Deduction A process where a general principle is applied to a particular case to explain it. For example, all humans die; this is a human, therefore he/she will die.

Deontological ethics A system of ethics which holds that some actions are, in principle, right, while others are wrong. According to this view, we must do the right thing, even when the consequences of our actions are not the best possible outcomes. For example, as health professionals we must do our best to save the life of an evil person even when we are certain that he will continue to do bad things.

Descriptive statistics Statistics designed to describe characteristics of a sample. For example, the most common or typical value or the extent of variation among such values.

Determinism The view that all events are caused by other events.

Directional hypothesis A directional hypothesis is one that asserts that differences between groups in the data will occur in a particular direction. For example, the hypothesis ‘smokers die younger than non-smokers’ is a directional hypothesis.

Discontinuous data Sometimes termed discrete data; variables that have discrete categories, e.g. male versus female.

Discussion A section of a research report in which the research findings are discussed.

Dispersion Sometimes known as variability the extent to which scores in a group of scores vary. This may be measured by statistics such as the standard deviation, variance, range and semi-interquartile range.

E

Ecological validity The extent to which the results of a study may be generalized to the real world.

Effect size The amount of change created by an intervention, especially in an experimental study.

Empathy In qualitative research, the ability to understand the perspectives of others.

Epidemiology The study of the distribution and determinants of disease within a community.

Ethics A project is ethical to the extent that its design and execution conform to a set of standards or conventions guiding research.

Ethics committees These are groups appointed by institutions, such as a university or a hospital, consisting of professional members such as researchers, ethicists and statisticians as well as people from the community. The role of the committee is to scrutinize the ethical and methodological standards of all research proposals which involve the institution. No research can be initiated until the committee decides to approve the project. It is essential that an explicit research protocol is presented to the committee, and that researchers scrupulously implement the protocol.

Ethnography A descriptive qualitative study, often of an individual or situation, usually written from the perspective of the participant(s) in the first person.

Ethnomethodology A qualitative approach to research which involves the study of social processes associated with the ways in which people perceive, describe and explain the world.

Evidence-based health care Defined as identifying and applying currently available and methodologically valid research evidence for selecting the best available practice for assessing or treating patients.

Expected frequency In the analysis of categorical data, the expected frequency is the one that would be expected in a particular category, under certain theoretical conditions. The expected frequency of women in a sample of 100 people would be 50, if equal proportions of sexes were assumed.

Experiment A research design involving the random allocation of subjects to groups and the application of different interventions to these groups. A non-intervention control group is often employed in an experimental

design. The aim of an experiment is to be able to conclude validly that differences in outcomes for the groups were caused by the different interventions.

Explanatory RCT An explanatory trial tests the *efficacy* of the intervention in a controlled research setting with selected participants and under ideal conditions. An efficacy study is designed to test whether the intervention works under ideal circumstances.

External validity The extent to which the results of the study may be generalized to the population.

Extreme response mode A method of responding to questions in which the respondent chooses the most extreme available response categories.

F

Factorial design A type of research design in which combinations of several independent variables are manipulated concurrently. A 2×2 factorial design involves the manipulation of two independent variables, each with two levels.

False negative The situation that occurs when a diagnostic test indicates that the person being assessed does not have a disease when he/she actually does.

False positive The situation that occurs when a diagnostic test indicates that the person being assessed has a disease when he/she does not.

Forced-response format A method of eliciting responses to a questionnaire in which there is no middle response category. This is sometimes done to avoid acquiescent response mode.

Forest plot A graph representing the outcomes of a meta-analysis. Typically, a forest plot shows the effect size and 95% confidence intervals for the individual trials and the overall or 'pooled' effect size.

Frequency distribution The way in which scores within a given sample or population are distributed.

Frequency polygon A method of graphing frequency distributions.

G

Grounded theory A qualitative research approach that advocates the development of theories to explain social phenomena grounded in data, following a process of induction, deduction and verification.

H

Hawthorne effect An effect which results in the improvement of subjects' performances through being observed and/or social contact. An example of a placebo effect.

Hermeneutics An approach to analysing the meaning of texts in the cultural contexts in which they were originally produced.

Histogram A method of graphing frequency distributions.

History A threat to the validity of studies in which unforeseen and uncontrolled events occur to the participants during the study that are outside the control of the researcher and which may be responsible for changes in the participants.

Hypothesis A proposition advanced by the researcher which is evaluated using the data collected.

I

Incidence rate The occurrence of new cases of a disease or condition within a specified time frame. See also Prevalence.

Incidental sample A method of sampling in which the researcher takes the most conveniently available cases.

Independent variable In an experiment an independent variable is the variable or condition manipulated by the researcher.

Induction The process in which a set of observations is made and a general principle formed to explain them. For example, ‘Every human I have read about eventually dies; this is a human, therefore I expect him/her to die’.

Informed consent The situation where a competent person, in possession of all the relevant facts, has agreed to participate in a research study.

Instrumentation In a study, instrumentation may be a threat to internal validity. It refers to the situation when the instrumentation changes over the period of the study, thus invalidating comparison of measured results.

Internal validity In a study, internal validity refers to the ability of the researcher to attribute differences in the groups or participants to the independent variable.

Inter-observer reliability The extent to which observers rating a particular phenomenon agree with each other.

Interrupted time series A type of research design in which a case is repeatedly measured over time to produce a series of measurements. The series is interrupted by an intervention or event, the effects of which may then be monitored by continuing the measurement series.

Interval scale A type of measurement scale with the following properties: (i) the values are distinguishable; (ii) they are ordered; (iii) the intervals between the points on the scale are equal; (iv) the zero point is not absolute, i.e. does not represent the absence of the quantity.

Interview A conversation between one or more interviewers and interviewees with the purpose of eliciting certain information.

Intra-observer reliability The extent to which an observer rating a particular phenomenon agrees with his/her own rating when presented with the same task on two different occasions.

L

Likert scales Likert scales are used in ‘closed response’ format questionnaires to quantify subjective beliefs and attitudes. Likert-like formats include a statement, e.g. ‘I love statistics’, followed by a five- or seven-point scale, with a neutral, middle choice; e.g. (1) strongly agree, (2) agree, (3) undecided, (4) disagree, (5) strongly disagree. We hope that you picked (1).

Literature review This is a section of a research report in which the previous research that has been done in the area is reviewed and related to the present problem being studied.

M

Matching In a study, subjects may be assigned to their groups using matching or random assignment. In matching in a two-group study, pairs of similar ‘matched’ subjects are formed and then one member of the pair is randomly assigned to one group and the other member to the other group. This ensures that the two groups have similar characteristics.

Maturation The phenomenon where participants in a study change spontaneously over time due to natural maturational changes. For example, children may grow older or an infection may spontaneously clear up.

Mean The average of a group of scores. For example, the mean of the scores 7, 8 and 9 is $(7 + 8 + 9) \div 3 = 8$. In statistical notation the mean of a sample is represented by the symbol $\bar{x}$. The mean of a population is represented by the symbol μ .

Measurement A procedure where qualities or quantities are attributed to characteristics of objects, persons or events. Weighing a patient involves a measurement process, as does a clinical judgement about whether a symptom is present or not.

Median The ‘middle’ score of a group of scores. For example, the median of the scores 7, 8 and 9 is 8. The median is the 50th percentile. The

median is often used in preference to the mean when a group of scores contains a small number of extremely small or large scores, because it is less sensitive to extreme values.

Meta-analysis Statistical approach to combining the results extracted from several related studies. The overall, pooled statistic represents the best estimate of the effect of a factor on health-related outcomes.

Mixed methods Refers to the combined use of both quantitative and qualitative methods for conducting a research project. The advantage of mixed methods is that the data produced by the study includes both objective facts and the subjective experiences of the participants.

Mode The mode is the most frequently occurring score in a group of scores. For example, the mode of the scores 7, 8, 8, 9, 10 is 8.

Mortality Used to describe a situation where some participants in a study are unable to continue in a study. This might be because they died or because they refuse to continue. If there is high mortality in a group in a study, this can jeopardize the internal validity of the study, because differences between the groups may be caused by differential mortality.

Multiple group time series A type of research design where two groups or cases are repeatedly measured over time to produce a series of measurements. One group or case receives an intervention and the other does not. The effects of the intervention may then be studied by comparing the two series.

N

n The symbol used to represent the number of cases in a sample.

N The symbol used to represent the number of cases in a population.

n = 1 design A research design in which one subject, rather than a group of subjects, is studied.

Narrative review A method of reviewing research literature that involves the narrative description of key articles selected and evaluated by the

researcher. Considered to be less rigorous than a systematic review.

Natural comparison study A type of study in which naturally occurring groups are compared with one another. For example, the health status of smokers versus non-smokers may be studied in a natural comparison study. The researcher does not assign the participants to the groups. These are naturally occurring. Studies of gender differences are natural comparisons.

Natural setting The normal setting of the phenomenon or people under study. Studies performed under laboratory conditions may sometimes have diminished external validity.

Negative correlation A correlation is a measure of the strength and direction of the association between two variables. A negative correlation between two variables implies that as one variable gets bigger, the value of the other variable becomes smaller.

Negative skew A frequency distribution where there is a long tail towards the negative end of the x-axis. [Fig. G.1](#) represents both positively and negatively skewed samples. The position of the 'tail' of the distribution determines whether the skew is positive or negative.

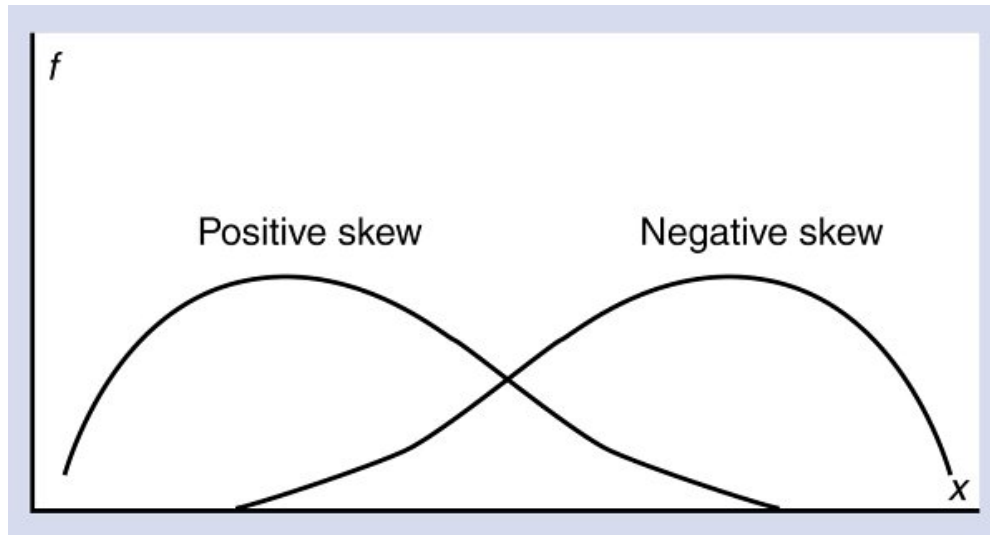


FIG. G.1 Positive and negative skew.

Nominal scale Measurement scales may be nominal, ordinal, interval or ratio. A nominal scale (often called a categorical scale) is one in which the values are distinct categories, e.g. male or female, Catholic or Protestant, Jewish or Muslim. It has the property of distinctiveness of values, but not ordering, equidistant intervals or an absolute zero.

Non-directional hypothesis This asserts that there are differences between groups in the data but with no direction specified. For example, the hypothesis ‘smokers and non-smokers have different life expectancies’ is a non-directional hypothesis.

Non-experimental study A study in which the researcher observes a situation but does not systematically manipulate or experiment with it. This may also be called a descriptive design.

Non-parametric test Statistical tests are chosen on the basis of the type of scales that are being analysed. Tests that are suitable for the analysis of ordinal or nominal data are termed non-parametric.

Normal curve A bell-shaped curve, as shown in [Fig. G.2](#). The normal curve is symmetrical and unimodal (has one peak).

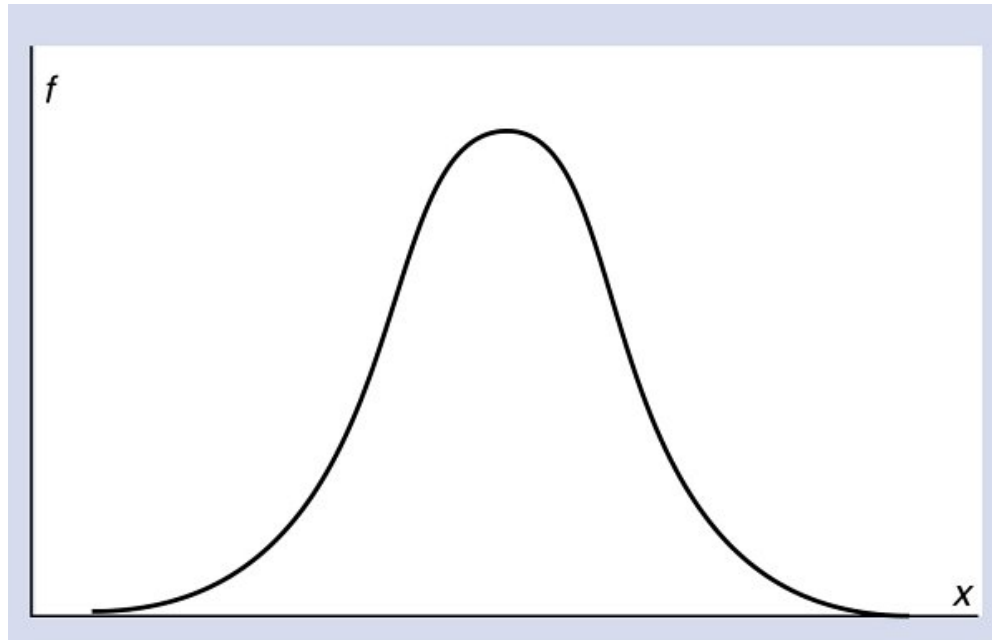


FIG. G.2 Normal curve.

Null hypothesis The hypothesis in a study that asserts there is no difference between groups or relationship between variables. The statistician normally poses the null hypothesis and then tests it statistically. If it is rejected, the alternative hypothesis (that there is a difference between two groups or a relationship between variables) is accepted. The null hypothesis is represented by the symbol H_0 .

O

Objective measures Measures derived from a mechanical measuring process involving a minimum amount of human interpretation, e.g. weight measurements.

Observation A situation where the researcher studies the phenomenon without deliberate intervention.

Observed frequency The actual number of occurrences of an event observed by a researcher. For example, the observed frequency of females who passed the exam in a group of students might be 37 students. See also Expected frequency.

Observer as participant Where the researcher studies the behaviour of a group by actively participating in the group's activities and situation.

Odds ratio (OR) An index of effect size suitable for nominal scale data. The OR provides a comparison of the magnitude risk or therapeutic benefits between an experimental and a control group.

One-tailed test A statistical test where a difference between two groups is expected to occur in a particular direction. For example, it may be hypothesized that smokers will have more health problems than non-smokers. This would be tested by applying a one-tailed test of significance.

Open-ended question If a question is asked without a predefined set of responses, it is an open-ended question. For example, the question 'What did you think about the program?' is an open-ended question.

Operational definition The specific way in which a concept or variable has been measured in a study. For example, the operational definition of anxiety in a study might be scores from the Spielberger State-Trait Anxiety Inventory or a self-rating on a 1–10 scale.

Ordinal scale Measurement scales may be nominal, ordinal, interval or ratio. An ordinal scale has the following properties: (i) the values are distinguishable; (ii) they are ordered but the intervals between the points are not equidistant, nor is there a meaningful zero point. The values of an ordinal scale are often ranked, e.g. 1st, 3rd.

P

Paradigm A general world view of health by the researcher. In the present book we argued for a 'post-positivist' paradigm, which includes the use of both quantitative and qualitative methods.

Parametric test Statistical tests are chosen on the basis of the type of data being analysed. Statistical tests that are suitable for the analysis of interval or ratio data are termed parametric.

Participation A situation where the researcher studies the phenomenon by actively participating.

Percentile The position of a score on a frequency distribution expressed in standard score units such that the 50th percentile is the median and all scores fall below the 100th percentile.

Phenomenology Phenomenology has its origins in both philosophy and psychology. In health science research, phenomenology is one of the methods used by qualitative researchers to study the way in which persons consciously experience health and illness in their lives. PICOT a rigorous method for specifying systematic review questions (P) Population refers to the participants in the studies (I) – Intervention refers to the treatments (C) – Comparison refers to what you will use as a reference group to compare with the intervention (O) – Outcome refers to what is measured (T) – Time describes the duration of the data collection.

Pie diagram A graphical method of representing the frequency distribution of a set of categorical data in the shape of a pie.

Pilot study A preliminary study where the procedures and protocols are tested or ‘piloted’.

Placebo effect The phenomenon where an otherwise worthless intervention in a study nevertheless induces an improvement in the patient’s condition or perception of the condition, perhaps due to the expectations of the participants in the study.

Population A group of people, institutions, cases or objects defined as that under study by the researcher. Samples are drawn from populations. Examples of populations are all coronary heart disease cases in the UK, all Australian men, and all ‘not-for-profit’ hospitals in Canada.

Population parameter A value derived from a population. For example, the average age of all the patients in a population defined by the

researcher is a population parameter. In statistical notation, population parameters are represented by Greek letters.

Pragmatic RCT An effectiveness or pragmatic RCT study tests the effectiveness of the intervention in everyday practice. The study includes participants representative of normal patients and with less controlled conditions that more closely reflect everyday clinical practice. Pragmatic or effectiveness RCTs are used in research translation studies to check that interventions that work under highly controlled circumstances will work in normal clinic circumstances.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA <http://www.prisma-statement.org>) is the global reporting standard for systematic reviews and meta-analyses

PROSPERO PROSPERO (International Prospective Register of Systematic Reviews) is an international systematic review registration system

Population validity The extent to which a sample reflects the characteristics of a population from which it is drawn.

Positive correlation A correlation is a measure of the strength and direction of the association between two variables. A positive correlation between two variables implies that as the values of one variable get bigger, so do the values of the other variable.

Positive skew A frequency distribution where there is a long tail towards the positive end of the x-axis.

Post-test only design A type of experimental study in which measurements of the groups are taken only after an intervention has occurred. This is generally done to avoid the effects of measurement upon people's response to the interventions.

Power The probability of rejecting the null hypothesis when the alternative hypothesis is true, i.e. correctly identifying an effect when it is there.

Pre-test/post-test design A type of experimental study in which measurements of the groups are taken both prior to and following an intervention. This allows the direct comparison of pre-intervention and post-intervention results for individual subjects and groups of subjects.

Predictive validity The extent to which a test or measure can validly predict a future event. For example, a clinical test may have high predictive validity with respect to 5-year mortality of people with cancer. It is often expressed in the form of a correlation coefficient.

Prevalence The overall occurrence of a particular disease in a specific population at a specific point in time.

Probability The chance or likelihood of an event. The probability of flipping a 'heads' with a fair coin is 0.5 or $\frac{1}{2}$. Probabilities may vary in value from 0 ('no chance') to 1 (certain).

Procedure A section in a research report that describes the protocol or procedure followed for the collection of data.

Proportion The ratio of one value to another expressed as a fraction of 1. For example, the proportion of women in the adult population is about 0.5.

Protocol deviation A deviation from the ideal procedure described as having been followed in a study by the researcher.

Q

QALY A QALY is the person's length of life multiplied by a valuation of their health-related quality of life. QALY measures are frequently used in the economic evaluation of health interventions.

QoL Quality of Life. The general well-being of populations and individuals normally measured by a standardised tool.

Qualitative methods An approach to research that emphasizes the non-numerical and interpretive analysis of social phenomena.

Quantitative methods An approach to research that emphasizes the collection of numerical data and the statistical analysis of hypotheses proposed by the researcher.

Quasi-experimental design A structured research design that is experiment-like but does not involve its full characteristics such as a control group and random assignment of subjects to treatment groups.

Questionnaire A means of collecting data from people where they provide written responses to a set of questions, either in their own words or by selecting predefined answers.

Quota sample The result of a sampling procedure in which the researcher sets quotas for the number of cases in particular categories to be included in the sample. For example, a sample of 100 might have quotas of 50 men and 50 women. The cases, however, are still selected on the basis of convenience rather than randomly. See also Stratified random sample.

R

Random assignment In an experiment, subjects are assigned to their groups by using a random assignment method. In random assignment, subjects are assigned to their groups using a random procedure. For example, in a two-group study the tossing of a coin to assign subjects to groups would be a random assignment procedure.

Random sample A group of cases drawn from a population such that each member of the population has had an equal chance of selection.

Random sampling The process of selecting cases from a population such that each member of the population has had an equal chance of selection.

Range In a group of scores the range is the difference between the maximum and minimum scores.

Ratio scale Measurement scales may be nominal, ordinal, interval or ratio. A ratio scale has the following properties: the values are distinctive, ordered and equidistant, and the zero point represents an absence of the

quantity rather than being an arbitrary zero. Metres and kilograms are examples of ratio scales.

RCT Randomised Controlled Trial (RCT). A research design aimed at assessing the effectiveness of a clinical intervention in which participants are randomly allocated to different treatment groups or a usual treatment, placebo or no treatment control group.

Refereed journal A journal in which the articles are vetted by independent referees for quality and interest. Refereed journals generally carry more highly regarded articles.

Regression to the mean The phenomenon where an individual who is measured on a test and obtains an extreme (very high or low) score and then upon re-measurement tends to move towards (regress) the average score (mean). Regression to the mean may be misinterpreted as representing a real change in score.

Relationships Associations between variables.

Reliability The extent to which a test or measurement result is reproducible.

Repeated measures The situation where a group of cases are measured on more than one occasion, for example prior to and following an intervention.

Representative sample A sample that accurately reflects the characteristics of the population from which it is drawn. Sometimes termed an 'unbiased' sample.

Reversal Reversal of the order of interventions is often used in experimental studies to control for the effects of order of administration. For example, intervention B preceded by intervention A may have a different effect than B followed by A.

Risk factors In epidemiology, a risk factor is an agent that is believed to increase the probability of a certain outcome or illness. For example, smoking may be a risk factor for the onset of coronary heart disease.

Rosenthal effect The phenomenon where the expectations of the researchers in a study influence the outcome. For example, if an observer believes that a particular intervention is effective, he/she may under-report or discount symptoms inconsistent with this belief.

S

Sample A group of cases selected from a population.

Sampling error Because a sample is smaller than the population from which it is drawn, there is often a discrepancy between the values obtained for the sample and those that apply to the population. For example, the average age of a sample might be 22 years and the average age of the population might be 28 years. This discrepancy is termed the sampling error.

Sampling method The method by which a sample is drawn from a population. Broadly, there are two approaches: random, in which every case in the population has an equal chance of selection; and non-random, in which cases have different chances of selection.

Scattergram A graph displaying the relationship between a set of paired scores on two variables. Also referred to as a scatterplot.

Semi-interquartile range A measure of the spread of a frequency distribution, being the distance between the first and third quartiles.

Semiotics A method of interpretation aiming to uncover the hidden, implicit meanings in texts.

Sensitivity This is a statistic used to analyse the accuracy of diagnostic or screening test for a disease. It represents the ratio of true positives and true negatives. A high degree of sensitivity indicates that the test is accurate in that it identifies the people who actually have the disease or condition. See also Specificity.

Shared variance In the examination of relationships between variables, researchers are concerned about the extent to which when one variable

changes the other also changes. Shared variance refers to the extent to which this occurs.

Skewness A characteristic of the shape of a frequency distribution. See also Negative skew, Positive skew.

Specificity This is a static relevant to analysing the accuracy of a diagnostic or screening tests for a disease. It represents the ratio of true negatives divided by the number of true negatives and false positives. A high degree of specificity indicates that the test is accurate; it correctly identifies those who do not have the disease or condition. See also Sensitivity.

Standard deviation A measure of the dispersion or variability of a group of scores. The standard deviation of a sample is represented by the symbol s and by the symbol σ for a population.

Standard error of the mean If a number of samples were taken from the same population and the mean calculated for each sample, the standard deviation of this distribution of means is known as the standard error of the mean. The standard error of the mean is used in statistical inferences about population means.

Standard normal curve An idealized normal or bell-shaped frequency distribution with a mean of 0 and a standard deviation of 1 unit.

Standardized test A standardized test is one that has known characteristics, especially known levels of reliability and validity. Researchers use standardized tests wherever possible in preference to unstandardized tests.

Statistic A statistic is a number with known properties derived from sample data. There are two types of statistics: inferential statistics, which are used to apply statistical tests; and descriptive statistics, which are used to describe characteristics of the sample.

Statistical significance When researchers have demonstrated statistical significance through the application of a statistical test, they have

demonstrated that the result obtained is probably not due to chance but is 'real'. This does not mean the result is important or interesting. See also Clinical significance.

Stratified random sample A type of sample in which the researcher wishes to ensure that important subgroups and their representation are preserved in the sample. For example, if a researcher randomly selected a subsample of 50 men and then 50 women, the sample has been stratified with respect to gender. See also Quota sample.

Structured interview An interview in which the questions are generally predefined, i.e. asked in a fixed order with the answers recorded by the researcher on a response sheet.

Subjective measures Measures derived from a measurement process involving a substantial degree of human interpretation, e.g. subjective ratings of pain, clinical ratings of social skills.

Subjects Participants in a study.

Surveys A type of research design in which characteristics of the cases under study are systematically recorded without the researcher attempting actively to change the situation. A non-experimental type of study.

Systematic Review A rigorous review of research evidence usually concerned with supporting a particular clinical intervention and generally based upon analysis of a group of Randomised Controlled Trial studies in the same content area.

Systematic sample A method of drawing a sample from a population where, for example, every 10th case is selected from a population list.

T

Test-retest reliability When a test or assessment procedure is administered twice to the same group of people, the correlation between

the first score and the second score is termed the test–retest reliability. This is a measure of the reproducibility of an assessment procedure.

Time series A series of measurements taken repeatedly from the same person, or group of people, over time.

Transcript A verbatim written version of an interview.

Transformed score A score that has been altered by an arithmetic manipulation, such as a z score.

True negative The situation that occurs when a diagnostic test indicates the person being assessed does not have a disease when he/she really does not.

True positive The situation that occurs when a diagnostic test indicates the person being assessed does have a disease when he/she does.

Two simple hyphentailed test A statistical test where a difference between two groups is tested without reference to the expected direction of the difference.

Type I error When researchers, on the basis of a statistical test applied to a sample of data, wrongly conclude that there is evidence of an association between variables or difference between groups in the population, they have committed a type I error. The probability of a type I error is represented by the symbol α .

Type II error A ‘miss’. When researchers, on the basis of a statistical test applied to a sample of data, wrongly conclude that there is no evidence of an association between variables or difference between groups in the population, they have committed a type II error. The probability of a type II error is represented by the symbol β .

U

Unstructured interview An interview in which there may be no pre-planned questions or fixed agenda. The dialogue is usually recorded in a transcript or field notes which are subsequently analysed.

Utilitarian ethics A form of consequentialism which holds that the morally correct action is the one which maximizes benefits for all the people involved. For example, immunizing children against diseases such as mumps may be mildly painful to a child, but it is the right action because it will protect the individual child and the community from the disease.

V

Validity The extent to which a test measures what it is intended to measure.

Variability The extent to which a group of scores varies or is spread out. This is usually measured by a descriptive statistic such as the standard deviation range or semi-interquartile range.

Variable A property or attribute that varies. For example, gender, age, weight are all variables.

Variance A measure of the dispersion or variability of a group of scores. The variance of a sample is represented by the symbol s^2 and the symbol σ^2 for a population. See also Standard deviation.

Z

z scores Transformed scores which express how many standard deviations a specific score is above or below the mean. For instance, a corresponding z score of 22 implies that a given score is two standard deviations below the mean.

Zero correlation Correlation is used to measure the strength and direction of an association between two variables. A zero correlation implies that two variables are unrelated to one another.

References and Further Reading

- Australian Commission on Safety and Quality in Health Care, . *Patient-Centred Care: Improving Quality and Safety Through Partnerships with Patients and Consumers* . Sydney: ACSQHC; 2011.
- Baggini J, Fosl P.S. *The Philosopher's Toolkit: A Compendium of Philosophical Concepts and Methods* . 2nd ed. Oxford: Wiley-Blackwell; 2010.
- Barker R.A, Barrett J, Mason S.L, Björklund A. Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson's disease. *Lancet Neurol* . 2013;12(1):84–91.
- Beaglehole R, Epping-Jordan J, Patel V, et al. Improving the prevention and management of chronic disease in low-income and middle-income countries: a priority for primary health care. *Lancet* . 2008;372(9642):940–949. doi: [10.1016/s0140-6736\(08\)61404-x](https://doi.org/10.1016/s0140-6736(08)61404-x).
- Beecher H.K. *Measurement of Subjective Responses: Quantitative Effects of Drugs* . Oxford: Oxford University Press; 1959.
- Björklund A, Lindvall O. Replacing dopamine neurons in Parkinson's disease: how did it happen? *J Parkinsons Dis* . 2017;7(s1):S21–S31.
- Bloom B.S. Effects of continuing medical education on improving physician clinical care and patient health: a review of systematic reviews. *Int J Technol Assess Health Care* . 2005;21(3):380–385.
- Bonita R, Beaglehole R, Kjellström T. *Basic Epidemiology* . 2nd ed. Geneva: World Health Organization; 2006.
- Bowling A. *Measuring Health: A Review of Subjective Health, Well-Being and Quality of Life Measurement Scales* . 4th ed. London: Open University Press; 2017.
- Browning C, Chapman A, Cowlshaw S, et al. The Happy Life Club™ study protocol: a cluster randomised controlled trial of a type 2 diabetes health coach intervention. *BMC Public Health* . 2011;11(1):90. doi: [10.1186/1471-2458-11-90](https://doi.org/10.1186/1471-2458-11-90).
- Browning C, Chapman A, Yang H, et al. Management of type 2 diabetes in China: the Happy Life Club, a pragmatic cluster randomised controlled

- trial using health coaches. *BMJ Open* . 2016;6(3):e009319. doi: [10.1136/bmjopen-2015-009319](https://doi.org/10.1136/bmjopen-2015-009319).
- Butler D. Translational research: crossing the valley of death. *Nature* . 2008;453(7197):840–842.
- Buttery P.C, Barker R.A. Treating Parkinson's disease in the 21st century: can stem cell transplantation compete? *J Comp Neurol* . 2014;522:2802–2816.
- Callard F, Rose D, Wykes T. Close to the bench as well as at the bedside: involving service users in all phases of translational research. *Health Expect* . 2011;15(4):389–400.
- Campbell S.M, Roland M.O, Buetow S.A. Defining quality of care. *Soc Sci Med* . 2000;51(11):1611–1625.
- Centre for Reviews and Dissemination, University of York. Prospero: International prospective register of systematic reviews. <https://www.crd.york.ac.uk/prospero/2017>.
- Chalmers A.F. *What is This Thing Called Science?* 4th ed. Brisbane: University of Queensland Press; 2013.
- Chalmers AF. *What Is This Thing Called Science?* St Lucia: University of Queensland Press.
- Chapman A, Liu S, Merkouris S, et al. Psychological interventions for the management of glycemic and psychological outcomes of type 2 diabetes mellitus in China: a systematic review and meta-analyses of randomized controlled trials. *Front Public Health* . 2015;3:252. doi: [10.3389/fpubh.2015.00252](https://doi.org/10.3389/fpubh.2015.00252).
- Coates S.J, Steed L, Price J. *SPSS: Analysis Without Anguish* . Sydney: John Wiley & Sons Australia Ltd.; 2003.
- Cochrane A.L. *Effectiveness and Efficiency: Random Reflections on Health Services* . London: Nuffield Trust; 1972.
- Cochrane Consumer Network. What is a systematic review? <http://consumers.cochrane.org/what-systematic-review>. Published 2017.
- Cohen J. *Statistical Power Analysis for the Behavioural Sciences* . 2nd ed. London: Routledge; 1988.

- Cohen J. Things I have learned (so far). *Am Psychol* . 1990;45(12):1304–1312.
- Cowlshaw S, Merkouris S, Dowling N, Anderson C.J, Jackson A.C, Thomas S.A. Psychological therapies for pathological and problem gambling. *Cochrane Database of Syst Rev* . 2012;11:CD008937. doi: [10.1002/14651858.CD008937.pub2](https://doi.org/10.1002/14651858.CD008937.pub2).
- Creswell J.W, Poth C.N. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* . 4th ed. Thousand Oaks, CA: Sage; 2018.
- Dal Grande E, Chittleborough C.R, Campostrini S, Taylor A.W. Bias of health estimates obtained from chronic disease and risk factor surveillance systems using telephone population surveys in Australia: results from a representative face-to-face survey in Australia from 2010 to 2013. *BMC Med Res Methodol* . 2016;16:44. doi: [10.1186/s12874-016-0145-z](https://doi.org/10.1186/s12874-016-0145-z).
- Daly J, Kellehear A, Glikman M. *The Public Health Researcher: A Methodological Guide* . Melbourne: Oxford University Press; 1997.
- Dawes M, Davies P, Gray A, Mant J, Seers K, Snowball R. *Evidence-Based Practice: A Primer for Health Care Professionals* . 2nd ed. Edinburgh: Churchill Livingstone; 2005.
- Denzin N.K, Lincoln Y.S. *The SAGE Handbook of Qualitative Research* . 5th ed. Thousand Oaks, CA: Sage; 2017.
- Dooley D. *Social Research Methods* . 4th ed. New Jersey: Prentice Hall; 2001.
- Doyle C, Lennox L, Bell D.A. Systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* . 2013;3(1):e001570 doi: [10.1136/bmjopen-2012-001570](https://doi.org/10.1136/bmjopen-2012-001570).
- Doyle J, Thomas S.A. Capturing policy in hearing-aid decisions by audiologists. In: Conolly T, Arkes H.R, Hammond K.R, eds. *Judgement and Decision Making: An Interdisciplinary Reader* . 2nd ed. Cambridge: Cambridge University Press; 2000:245–258.
- Duncan P. *Values, Ethics and Healthcare* . London: Sage; 2010.

- Ellis P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* . Cambridge: Cambridge University Press; 2010.
- Emanuel E.J, Wendler D, Grady C. What makes clinical research ethical? *JAMA* . 2000;283(20):2701–2711.
- Engel. 2003.
- Engel G.L. The clinical application of the biopsychosocial model. In: Frankel R, Quill T, McDaniel S, eds. *The Biopsychosocial Approach: Past, Present, Future* . New York, NY: University of Rochester Press; 2009.
- Epstein L, Ogden J. A qualitative study of GPs' views of treating obesity. *Br J Gen Pract* . 2005;55(519):750–754.
- Evans W.P. Breast cancer screening: successes and challenges. *CA Cancer J Clin* . 2012;62(1):5–9.
- Evidence-Based Working Group. 1992.
- Fara P. *Science: A Four Thousand Year History* . New York, NY: Oxford University Press; 2009.
- Florey C.D. Sample size for beginners. *BMJ* . 1993;306(6886):1181–1184. doi: [10.1136/bmj.306.6886.1181](https://doi.org/10.1136/bmj.306.6886.1181).
- Gerrig R.J, Zimbardo P.G, Campbell A.J, et al. *Psychology and Life* . 2nd Australasian ed. Frenchs Forest, NSW, Australia: Pearson; 2012.
- Gibson D.G, Pereira A, Farrenkopf B.A, Labrique A.B, Pariyo G.W, Hyder A.A. Mobile phone surveys for collecting population-level estimates in low- and middle-income countries: a literature review. *J Med Internet Res* . 2010;19(5):e139. doi: [10.2196/jmir.7428](https://doi.org/10.2196/jmir.7428).
- Glass G.V. Primary, secondary, and meta-analysis of research. *Educ Res* . 1976;5(10):3–8. doi: [10.3102/0013189X005010003](https://doi.org/10.3102/0013189X005010003).
- Guba E.G, Lincoln Y.S. Epistemological and methodological bases of naturalistic inquiry. *EC & TJ* . 1982;30(4):233–252.
- Guba E.G, Lincoln Y.S. Competing paradigms in qualitative research. In: Denzin N.K, Lincoln Y.S, eds. *Handbook of Qualitative Research* . London: Sage; 1994:105–117.

- Han X, Williams S.R, Zuckerman B.L. A snapshot of translational research funded by the National Institutes of Health (NIH): a case study using behavioral and social science research awards and Clinical and Translational Science Awards funded publications. *PloS One* . 2018;13(5):e0196545. doi: [10.1371/journal.pone.0196545](https://doi.org/10.1371/journal.pone.0196545).
- Hay D, Oken D. The psychological stress of intensive care unit nursing. In: Monat A, Lazarus RS, eds. *Stress and Coping*. 1991. New York: Columbia University Press; 118–131.
- Hicks C.M. Applied project design and analysis. In: Hicks C.M, ed. *Research Methods for Clinical Therapists: Applied Project Design and Analysis* . 5th ed. London: Elsevier Health Sciences; 2009.
- Higgins J.P.T, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0* . Melbourne: The Cochrane Collaboration; 2011.
- Hjörleifsson S, Hammer E, Díaz E. General practitioners' strategies in consultations with immigrants in Norway-practice-based shared reflections among participants in focus groups. *Fam Pract* . 2017;35(2):216–221. doi: [10.1093/fampra/cmz097](https://doi.org/10.1093/fampra/cmz097).
- Hossain P, Kavar B, El Nahas M. Obesity and diabetes in the developing world—a growing challenge. *N Engl J Med* . 2007;356(1):213–215.
- Huck S.W, Cormier W.H, Bounds W.G. *Reading Statistics and Research* . New York: Harper & Row; 1974.
- Iansek R, Morris M.E, eds. *Rehabilitation in Movement Disorders* . Cambridge: Cambridge University Press; 2013 doi: [10.1017/CBO9781139012942](https://doi.org/10.1017/CBO9781139012942).
- Jiang F, Zhang J, Shen X. Towards evidence-based public health policy in China. *Lancet* . 2013;381(9882):1962–1964. doi: [10.1016/S0140-6736\(13\)61083-1](https://doi.org/10.1016/S0140-6736(13)61083-1).
- Jones M.P, Attia J.R. Sampling: how you choose people is as important as how you analyse their data. *Med J Aust* . 2017;206(2):67–68.
- Khan K.S, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. *J R Soc Med* . 2003;96(3):118–121.

- Khoury M.J, Gwinn M, Yoon P.W, Dowling N, Moore C.A, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med* . 2007;9(10):665–674.
- Kindon S, Pain R, Kesby M. *Participatory Action Research Approaches and Methods: Connecting People*. Routledge: New York.
- Laing RD, Esterson A. *Sanity, Madness and the Family*. London: Penguin. 1990.
- Laue J, Melbye H, Halvorsen P.A, et al. How do general practitioners implement decision-making regarding COPD patients with exacerbations? An international focus group study. *Int J Chron Obstruct Pulmon Dis* . 2016;11:3109–3119. doi: [10.2147/COPD.S118856](https://doi.org/10.2147/COPD.S118856).
- Leung F.-H, Savithiri R. Spotlight on focus groups. *Can Fam Physician* . 2009;55(2):218–219.
- Li H.C, Chan S.S, Lam T.H. Smoking among Hong Kong Chinese women: behavior, attitudes and experience. *BMC Public Health* . 2015;15:183. doi: [10.1186/s12889-015-1529-4](https://doi.org/10.1186/s12889-015-1529-4).
- Liamputtong P. Qualitative data analysis: conceptual and practical considerations. *Health Promot J Aust* . 2009;20(2):133–139.
- Liamputtong P. *Qualitative Research Methods* . 4rd ed. Melbourne: Oxford University Press; 2009.
- Lincoln Y.S, Guba E.G. *Naturalistic Inquiry* . London: Sage; 1985.
- Lorains F.K, Cowlshaw S, Thomas S.A. Prevalence of comorbid disorders in problem and pathological gambling: systematic review and meta-analysis of population surveys. *Addiction* . 2011;106(3):490–498. doi: [10.1111/j.1360-0443.2010.03300.x](https://doi.org/10.1111/j.1360-0443.2010.03300.x).
- Lu A.P, Jia H.W, Xiao C, Lu Q.P. Theory of traditional Chinese medicine and therapeutic method of diseases. *World J Gastroenterol* . 2004;10(13):1854–1856.
- Masic I, Miokovic M, Muhamedagic B. Evidence based medicine—new approaches and challenges. *Acta Inform Med* . 2008;16(4):219–225.
- McDowell I. *Measuring Health: A Guide to Rating Scales and Questionnaires* . 3rd ed. New York: Oxford University Press; 2006.

- McGartland M, Polgar S. Paradigm collapse in psychology: the necessity for a 'two methods' approach. *Aust Psychol* . 1994;29(1):21–28.
- McGinn F. The plight of rural parents caring for adult children with HIV. *Fam Soc* . 1996;77(5):269–278.
- McQuarrie E.F, McIntyre S.H. Focus Groups and the Development of New Products by Technologically Driven Companies: some guidelines. *J Prod Innov Manag* . 1986;vol. 3(Issue1):40–47 March 1986.
- Miles M.B, Huberman A.M. *Qualitative Data Analysis: A Sourcebook for New Methods* . Beverly Hills, California: Sage; 1984.
- Minichiello V, Aroni R, Hays T. *In-Depth Interviewing* . 3rd ed. Sydney: Pearson Education Australia; 2008.
- Morgan DL, Morgan RK. *Single-Case Research Methods for the Behavioral and Health Sciences*. 2008. Los Angeles: SAGE.
- Morse J.M, Field P.A. *Nursing Research: The Application of Qualitative Approaches* . 2nd ed. Cheltenham: Nelson Thornes; 2002.
- Morse. *Field* . 2003.
- Mostofsky E, Burger M, Schlaug G. Alcohol and acute ischemic stroke onset: the stroke onset study. *Stroke* . 2010;41(9):1845–1849.
- Muir Gray J.A. *Evidence-Based Healthcare and Public Health: How to Make Decisions About Health Services and Public Health* . 3rd ed. London: Churchill Livingstone; 2009.
- Ng M, Fleming T, Robinson M, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* . 2014;384(9945):766–781.
- Okasha S. *Philosophy of Science: Very Short Introduction* . 2nd ed. Oxford: Oxford University Press; 2016.
- Olanow C.W, Goetz C.G, Kordower J.H, et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Ann Neurol* . 2003;54(3):403–414.
- Olanow C.W. The scientific basis for the current treatment of Parkinson's disease. *Annu Rev Med* . 2004;55:41–60.

- Ostan R, Monti D, Guerresi P, Bussolotto M, Franceschi C, Baggio G. Gender, aging and longevity in humans: an update of an intriguing/neglected scenario paving the way to a gender-specific medicine. *Clin Sci (Lond)*. 2016;130(19):1711–1725. doi: [10.1042/CS20160004](https://doi.org/10.1042/CS20160004).
- Pallant J. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*. Melbourne: Allen and Unwin; 2016.
- Patton M.Q. *Qualitative Evaluation and Research Methods*. 2nd ed. Newbury Park, CA: Sage; 1990.
- Polgar S, Karimi L, Buultjens M, Morris M.E. A critical evaluation of the methodological obstacles to translating cell-based research into an effective treatment for people with Parkinson's disease. *Neurorehab Neural Repair*. 2016;30(9):845–853.
- Polgar S, Morris M.E, Reilly S, Bilney B, Sanberg P.R. Reconstructive neurosurgery for Parkinson's disease: a systematic review and preliminary meta-analysis. *Brain Res Bull*. 2003;60(1-2):1–24.
- Polgar S, Ng J. Ethics, methodology and the use of placebo controls in surgical trials. *Brain Res Bull*. 2005;67(4):290–297.
- Polgar S, Swerissen H. *Research Methods in Health 2. Bundoora*. Victoria, Australia: La Trobe University; 2000.
- Polgar S, Thomas S.A. *Introduction to Research in the Health Sciences*. 7th ed. London: Elsevier Scientific; 2019.
- Raftery J, Young A, Stanton L, et al. Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme. *Health Technol Assess*. 2015;19(11):1–138. doi: [10.3310/hta19110](https://doi.org/10.3310/hta19110).
- Ren Y, Yang H, Browning C, Thomas S, Liu M. Therapeutic effects of motivational interviewing on blood pressure control: a meta-analysis of randomized controlled trials. *Int J Cardiol*. 2014;172(2):509–511. doi: [10.1016/j.ijcard.2014.01.051](https://doi.org/10.1016/j.ijcard.2014.01.051).
- Rice P.L, Ezzy D. *Qualitative Research Methods: A Health Focus*. Oxford: Oxford University Press; 1999.

- Riva J.J, Malik K.M, Burnie S.J, Endicott A.R, Busse J.W. What is your research question? An introduction to the PICOT format for clinicians. *J Can Chiropr Assoc* . 2012;56(3):167–171.
- Roddiss JK, Liversedge HL, Ryder I, Woodhouse M. Incorporating the patient experience into clinical guidelines: recommendations for researchers and guideline developers. *BMJ Evid Based Med*. Published Online: September 18, 2018. <https://doi.org/10.1136/bmjebm-2018-111015>.
- Rosenhan DL. On being sane in insane places. In: Krupat E, ed. *Psychology Is Social: Readings and Conversations in Social Psychology*. 1973. Glenview, Illinois: Scott Foresman;189-200.
- Rubio D.M, Schoenbaum E.E, Lee L.S, et al. Defining translational research: implications for training. *Acad Med* . 2010;85(3):470–475.
- Sackett D.L, Strauss S.E, Richardson W.S, Rosenberg W, Haynes R.B. *Evidence-Based Medicine: How to Practice and Teach EBM* . 2nd ed. Edinburgh: Churchill Livingstone; 2000.
- Schwartz M, Polgar S. *Statistics for Evidence-Based Health Care* . Croydon, Australia: Tertiary Press; 2003.
- Schiøtz M.L, Høst D, Christensen M.B, et al. Quality of care for people with multimorbidity – a case series. *BMC Health Serv Res* . 2017;17(1):745. doi: [10.1186/s12913-017-2724-z](https://doi.org/10.1186/s12913-017-2724-z).
- Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* . 2015;350:g7647. doi: [10.1136/bmj.g7647](https://doi.org/10.1136/bmj.g7647).
- Shin S.S, Xiao D, Cao M, et al. Patient and doctor perspectives on incorporating smoking cessation into tuberculosis care in Beijing, China. *Int J Tuberc Lung Dis* . 2012;16(1):126–131. doi: [10.5588/ijtld.11.0044](https://doi.org/10.5588/ijtld.11.0044).
- Skjeie H, Brekke M. ‘Big needles, small bodies’–the absence of acupuncture treatment for infants in contemporary Shanghai: a qualitative study. *BMJ Open* . 2015;5(11):e009486. doi: [10.1136/bmjopen-2015-009486](https://doi.org/10.1136/bmjopen-2015-009486).

- Smith C.M. Origin and uses of *primum non nocere*—above all, do no harm!. *J Clin Pharmacol* . 2013;45(4):371–377.
- Soares-Weiser K, Thomas S, Thomson G, Garner P. Ribavirin for Crimean-Congo hemorrhagic fever: systematic review and meta-analysis. *BMC Infect Dis* . 2010;10:207.
- Starks H, Trinidad S.B. Choose your method: a comparison of phenomenology, discourse analysis, and grounded theory. *Qual Health Res* . 2007;17(10):1372–1380.
- Strauss A, Corbin J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* . 2nd ed. Newbury Park, CA: Sage; 1990.
- Streiner D.L, Norman G.R, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use* . 5th ed. New York: Oxford University Press; 2015.
- Sullivan G.M, Feinn R. Using effect size—or why the P value is not enough. *J Grad Med Educ* . 2012;4(3):279–282. doi: [10.4300/JGME-D-12-00156.1](https://doi.org/10.4300/JGME-D-12-00156.1).
- Tashakkori A.M, Teddlie C.B. *Handbook of Mixed Methods in Social and Behavioral Research* . Thousand Oaks, CA: Sage; 2003.
- Taylor R. *Medicine Out of Control: The Anatomy of Malignant Technology* . Melbourne: Sun Books; 1979.
- Teltscher B, Polgar S. Objective knowledge about Huntington’s disease and attitudes towards predictive tests of persons at risk. *J Med Genet* . 1981;18(1):31–39.
- Thomas S.A, Browning C.J. Burden of disease. In: Pachana N, ed. *Encyclopedia of Geropsychology* . Berlin: Springer-Verlag; 2016.
- Thomas S.A, Merkouris S.S, Browning C.J, et al. The PROblem Gambling RESearch Study (PROGRESS) research protocol: a pragmatic randomised controlled trial of psychological interventions for problem gambling. *BMJ Open* . 2015;5:e009385 doi: [10.1136/bmjopen-2015-009385](https://doi.org/10.1136/bmjopen-2015-009385).
- Thomas S.A, Steven I, Browning C, et al. Focus groups in health research: a methodological review. *Ann Rev Hlth Soc Sci* . 1992;2(1):7–20.

- Thomas S, Steven I, Browning C, et al. Patient knowledge, opinions, satisfaction and choices in primary health care provision: a progress report. In: Dossel D.P, ed. *The General Practice Evaluation Program: The 1992 Work-in-Progress Conference* . Canberra: Australian Government Publishing Service; 1993.
- van Gelder M.M, Bretveld R.W, Roeleveld N. Web-based questionnaires: the future in epidemiology? *Am J Epidemiol* . 2010;172(11):1292–1298. doi: [10.1093/aje/kwq291](https://doi.org/10.1093/aje/kwq291).
- Van Lith T, Fenner P, Schofield P. The lived experience of art making as a companion to the mental health recovery process. *Disabil Rehabil* . 2011;33:652–660.
- Walker Q.J, Langlands A.O. The misuse of mammography in the management of breast cancer. *Med J Aust* . 1986;145(5):185–187.
- Woolf S.H. The meaning of translational research and why it matters. *JAMA* . 2008;299(2):211–213.
- World Health Organization. Traditional medicine strategy 2002–2005. World Health Organization: Geneva.
- World Health Organization. *Regional Guidelines for Continuing Medical Education (CME)/Continuing Professional Development (CPD) Activities* . New Delhi: World Health Organization Regional Office for South East Asia; 2010.
- World Medical Association. *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects* . Edinburgh: 52nd WMA General Assembly; 2008.
- Yung K.P, Chung K.F, Ho F.Y, Yeung W.F, Ng T.H. The experience of chronic insomnia in Chinese adults: a study using focus groups and insomnia experience diaries. *Behav Sleep Med* . 2016;14(4):406–428. doi: [10.1080/15402002.2015.1017097](https://doi.org/10.1080/15402002.2015.1017097).
- Zhao Y, Chen R, Wang B, Wu T, Huang Y, Guo A. General practice on-the-job training in Chinese urban community: a qualitative study on needs and challenges. *PLoS One* . 2014;9(4):e94301. doi: [10.1371/journal.pone.0094301](https://doi.org/10.1371/journal.pone.0094301).

A APPENDIX

z -Scores and Associated Areas Between z and Mean and Beyond

z	Area between mean and z	Area beyond z
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

z	Area between mean and z	Area beyond z
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
0.41	0.1591	0.3409
0.42	0.1628	0.3372

z	Area between mean and z	Area beyond z
0.43	0.1664	0.3336
0.44	0.1700	0.3300
0.45	0.1736	0.3264
0.46	0.1772	0.3228
0.47	0.1808	0.3192
0.48	0.1844	0.3156
0.49	0.1879	0.3121
0.50	0.1915	0.3085
0.51	0.1950	0.3050
0.52	0.1985	0.3015
0.53	0.2019	0.2981
0.54	0.2054	0.2946
0.55	0.2088	0.2912
0.56	0.2123	0.2877
0.57	0.2157	0.2843
0.58	0.2190	0.2810
0.59	0.2224	0.2776
0.60	0.2257	0.2743
0.61	0.2291	0.2709
0.62	0.2324	0.2676
0.63	0.2357	0.2643
0.64	0.2389	0.2611

z	Area between mean and z	Area beyond z
0.65	0.2422	0.2578
0.66	0.2454	0.2546
0.67	0.2486	0.2514
0.68	0.2517	0.2483
0.69	0.2549	0.2451
0.70	0.2580	0.2420
0.71	0.2611	0.2389
0.72	0.2642	0.2358
0.73	0.2673	0.2327
0.74	0.2704	0.2296
0.75	0.2734	0.2266
0.76	0.2764	0.2236
0.77	0.2794	0.2206
0.78	0.2823	0.2177
0.79	0.2852	0.2148
0.80	0.2881	0.2119
0.81	0.2910	0.2090
0.82	0.2939	0.2061
0.83	0.2967	0.2033
0.84	0.2995	0.2005
0.85	0.3023	0.1977
0.86	0.3051	0.1949

z	Area between mean and z	Area beyond z
0.87	0.3078	0.1922
0.88	0.3106	0.1894
0.89	0.3133	0.1867
0.90	0.3159	0.1841
0.91	0.3186	0.1814
0.92	0.3212	0.1788
0.93	0.3238	0.1762
0.94	0.3264	0.1736
0.95	0.3289	0.1711
0.96	0.3315	0.1685
0.97	0.3340	0.1660
0.98	0.3365	0.1635
0.99	0.3389	0.1611
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
1.04	0.3508	0.1492
1.05	0.3531	0.1469
1.06	0.3554	0.1446
1.07	0.3577	0.1423
1.08	0.3599	0.1401

z	Area between mean and z	Area beyond z
1.09	0.3621	0.1379
1.10	0.3643	0.1357
1.11	0.3665	0.1335
1.12	0.3686	0.1314
1.13	0.3708	0.1292
1.14	0.3729	0.1271
1.15	0.3749	0.1251
1.16	0.3770	0.1230
1.17	0.3790	0.1210
1.18	0.3810	0.1190
1.19	0.3830	0.1170
1.20	0.3849	0.1151
1.21	0.3869	0.1131
1.22	0.3888	0.1112
1.23	0.3907	0.1093
1.24	0.3925	0.1075
1.25	0.3944	0.1056
1.26	0.3962	0.1038
1.27	0.3980	0.1020
1.28	0.3997	0.1003
1.29	0.4015	0.0985
1.30	0.4032	0.0968

z	Area between mean and z	Area beyond z
1.31	0.4049	0.0951
1.32	0.4066	0.0934
1.33	0.4082	0.0918
1.34	0.4099	0.0901
1.35	0.4115	0.0885
1.36	0.4131	0.0869
1.37	0.4147	0.0853
1.38	0.4162	0.0838
1.39	0.4177	0.0823
1.40	0.4192	0.0808
1.41	0.4207	0.0793
1.42	0.4222	0.0778
1.43	0.4236	0.0764
1.44	0.4251	0.0749
1.45	0.4265	0.0735
1.46	0.4279	0.0721
1.47	0.4292	0.0708
1.48	0.4306	0.0694
1.49	0.4319	0.0681
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643

z	Area between mean and z	Area beyond z
1.53	0.4370	0.0630
1.54	0.4382	0.0618
1.55	0.4394	0.0606
1.56	0.4406	0.0594
1.57	0.4418	0.0582
1.58	0.4429	0.0571
1.59	0.4441	0.0559
1.60	0.4452	0.0548
1.61	0.4463	0.0537
1.62	0.4474	0.0526
1.63	0.4484	0.0516
1.64	0.4495	0.0505
1.65	0.4505	0.0495
1.66	0.4515	0.0485
1.67	0.4525	0.0475
1.68	0.4535	0.0465
1.69	0.4545	0.0455
1.70	0.4554	0.0446
1.71	0.4564	0.0436
1.72	0.4573	0.0427
1.73	0.4582	0.0418
1.74	0.4591	0.0409

z	Area between mean and z	Area beyond z
1.75	0.4599	0.0401
1.76	0.4608	0.0392
1.77	0.4616	0.0384
1.78	0.4625	0.0375
1.79	0.4633	0.0367
1.80	0.4641	0.0359
1.81	0.4649	0.0351
1.82	0.4656	0.0344
1.83	0.4664	0.0336
1.84	0.4671	0.0329
1.85	0.4678	0.0322
1.86	0.4686	0.0314
1.87	0.4693	0.0307
1.88	0.4699	0.0301
1.89	0.4706	0.0294
1.90	0.4713	0.0287
1.91	0.4719	0.0281
1.92	0.4726	0.0274
1.93	0.4732	0.0268
1.94	0.4738	0.0262
1.95	0.4744	0.0256
1.96	0.4750	0.0250

z	Area between mean and z	Area beyond z
1.97	0.4756	0.0244
1.98	0.4761	0.0239
1.99	0.4767	0.0233
2.00	0.4772	0.0228
2.01	0.4778	0.0222
2.02	0.4783	0.0217
2.03	0.4788	0.0212
2.04	0.4793	0.0207
2.05	0.4798	0.0202
2.06	0.4803	0.0197
2.07	0.4808	0.0192
2.08	0.4812	0.0188
2.09	0.4817	0.0183
2.10	0.4821	0.0179
2.11	0.4826	0.0174
2.12	0.4830	0.0170
2.13	0.4834	0.0166
2.14	0.4838	0.0162
2.15	0.4842	0.0158
2.16	0.4846	0.0154
2.17	0.4850	0.0150
2.18	0.4854	0.0146

z	Area between mean and z	Area beyond z
2.19	0.4857	0.0143
2.20	0.4861	0.0139
2.21	0.4864	0.0136
2.22	0.4868	0.0132
2.23	0.4871	0.0129
2.24	0.4875	0.0125
2.25	0.4878	0.0122
2.26	0.4881	0.0119
2.27	0.4884	0.0116
2.28	0.4887	0.0113
2.29	0.4890	0.0110
2.30	0.4893	0.0107
2.31	0.4896	0.0104
2.32	0.4898	0.0102
2.33	0.4901	0.0099
2.34	0.4904	0.0096
2.35	0.4906	0.0094
2.36	0.4909	0.0091
2.37	0.4911	0.0089
2.38	0.4913	0.0087
2.39	0.4916	0.0084
2.40	0.4918	0.0082

z	Area between mean and z	Area beyond z
2.41	0.4920	0.0080
2.42	0.4922	0.0078
2.43	0.4925	0.0075
2.44	0.4927	0.0073
2.45	0.4929	0.0071
2.46	0.4931	0.0069
2.47	0.4932	0.0068
2.48	0.4934	0.0066
2.49	0.4936	0.0064
2.50	0.4938	0.0062
2.51	0.4940	0.0060
2.52	0.4941	0.0059
2.53	0.4943	0.0057
2.54	0.4945	0.0055
2.55	0.4946	0.0054
2.56	0.4948	0.0052
2.57	0.4949	0.0051
2.58	0.4951	0.0049
2.59	0.4952	0.0048
2.60	0.4953	0.0047
2.61	0.4955	0.0045
2.62	0.4956	0.0044

z	Area between mean and z	Area beyond z
2.63	0.4957	0.0043
2.64	0.4959	0.0041
2.65	0.4960	0.0040
2.66	0.4961	0.0039
2.67	0.4962	0.0038
2.68	0.4963	0.0037
2.69	0.4964	0.0036
2.70	0.4965	0.0035
2.71	0.4966	0.0034
2.72	0.4967	0.0033
2.73	0.4968	0.0032
2.74	0.4969	0.0031
2.75	0.4970	0.0030
2.76	0.4971	0.0029
2.77	0.4972	0.0028
2.78	0.4973	0.0027
2.79	0.4974	0.0026
2.80	0.4974	0.0026
2.81	0.4975	0.0025
2.82	0.4976	0.0024
2.83	0.4977	0.0023
2.84	0.4977	0.0023

z	Area between mean and z	Area beyond z
2.85	0.4978	0.0022
2.86	0.4979	0.0021
2.87	0.4979	0.0021
2.88	0.4980	0.0020
2.89	0.4981	0.0019
2.90	0.4981	0.0019
2.91	0.4982	0.0018
2.92	0.4982	0.0018
2.93	0.4983	0.0017
2.94	0.4984	0.0016
2.95	0.4984	0.0016
2.96	0.4985	0.0015
2.97	0.4985	0.0015
2.98	0.4986	0.0014
2.99	0.4986	0.0014
3.00	0.4987	0.0013
3.01	0.4987	0.0013
3.02	0.4987	0.0013
3.03	0.4988	0.0012
3.04	0.4988	0.0012
3.05	0.4989	0.0011
3.06	0.4989	0.0011

z	Area between mean and z	Area beyond z
3.07	0.4989	0.0011
3.08	0.4990	0.0010
3.09	0.4990	0.0010
3.10	0.4990	0.0010
3.11	0.4991	0.0009
3.12	0.4991	0.0009
3.13	0.4991	0.0009
3.14	0.4992	0.0008
3.15	0.4992	0.0008
3.16	0.4992	0.0008
3.17	0.4992	0.0008
3.18	0.4993	0.0007
3.19	0.4993	0.0007
3.20	0.4993	0.0007
3.21	0.4993	0.0007
3.22	0.4994	0.0006
3.23	0.4994	0.0006
3.24	0.4994	0.0006
3.25	0.4994	0.0006
3.30	0.4995	0.0005
3.35	0.4996	0.0004
3.40	0.4997	0.0003

z	Area between mean and z	Area beyond z
3.45	0.4997	0.0003
3.50	0.4998	0.0002
3.60	0.4998	0.0002
3.70	0.4999	0.0001
3.80	0.4999	0.0001
3.90	0.49995	0.00005
4.00	0.49997	0.00003

APPENDIX B

t Distribution

Directional <i>p</i>	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
Non-directional <i>p</i>	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
Degrees of freedom								
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.213
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090

APPENDIX C

Chi-square (χ^2)

df	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.00016	0.00063	0.0039	0.016	0.064	0.15	0.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	0.02	0.04	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	0.12	0.18	0.35	0.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	0.30	0.43	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	0.55	0.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	0.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
df	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
22	9.54	10.60	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

Index

Note: Page numbers followed by “f” indicate figures, “t” indicate tables and “b” indicate boxes.

A

Absolute risk reduction, [144](#)

Absolute risk reduction (ARR), [144](#)

Abstract, [15](#)

Acquiescent response mode, [68](#)

Aims of research, *See* [Research](#)

α (alpha) decision level, [133](#)

‘Alternative hypothesis’, [133](#)

Analytical epidemiology, [51–52](#)

Anonymity, [36](#)

Apparatus and tools (research), [15](#)

 critical evaluation of published research, [161](#)

Appendices, research publications, [16](#)

Appraisal of published research, *See* [Critical evaluation of published research](#)

Area sampling, [29](#)

Area under normal curve, [110](#) , [110f–111f](#)

Assignment, [161–162](#)

critical evaluation of published research, [161–162](#)

independent groups, [43–44](#) , [45f](#)

matched groups, [44](#)

of participants into groups, [43–44](#)

randomized-controlled trial (RCT), [46](#)

Audiotaping, [78](#)

Auditability, [155](#)

Average deviation, [105](#) , [105t](#)

B

Bar graph, [95](#) , [96f](#)

Basic research, [177](#)

‘Bedside’, defined, [179](#)

‘Before-after’ design, [42](#) , [42f](#)

Bell-shaped curve, [109](#) , [109f](#)

Binomial theorem, [126](#)

Bioethics, [34](#)

Biopsychosocial model, [7](#)

Bracketing, [59](#)

C

Case-control designs, [49](#) , [52](#)

Case study, [49](#)

Causality

- concept of, 42
 - criteria, 42
 - correlation and, 119–120
- Census, 51
- Census studies, 26
- Central limit theorem, 126
- Central tendency, 102–107 , 105f
 - mean, 102–103
 - median, 103–104 , 103f , 103t
 - mode, 103–104 , 103t–104t
- Chain sampling, 31–32 , 32t
- Chi-square test (χ^2) test, 136–137 , 137t
 - appropriate, 136
 - contingency tables and, *See* Contingency tables
- Clinical decision making, 147–148
 - type I error, 145f , 147–148 , 148t
 - type II error, 147–148 , 148f , 148t
- Clinical research, 177
 - phases of, 180–181
- Clinical significance, 141
- Closed-response question formats, 67–68 , 67t
- Closing statements, 69
- Cochrane Collaboration
 - evidence-based practice, 175
 - systematic reviews, 171

Coding, of data, 82 , 151–152

predetermined coding, 151–152

thematic analysis, 152

units of analysis, 152

concepts, 152

sentences, 152

themes, 152

words, 152

Coefficient of determination, 119

Cognitive problems, relating to people with, 58

Cohen's *d*, 144

Cohort study, 49 , 53

Collection methods, recording information, 78–80 , 79t

Computer-assisted data analysis, 153

meta-analysis, 173

Confidence intervals, 121–131

calculating, 128 , 129f

effect sizes and, 144–145

in health research, 130–131

meta-analysis interpretation of results, 173

statistical significance and, 144–145 , 145f

using *t* distribution, 128–130 , 129f–130f

‘Confidentiality’, 36

Confounding and bias, 42–43

‘before-after’ design, 42 , 42f

history of, [42](#)

internal validity, [43](#)

maturation, [42](#)

observer bias, [42](#)

placebo effects, [42](#)

Consent, informed, *See* [Informed consent](#)

Construct validity, [87](#)

Content analysis, [152–153](#)

example, [153](#)

Content validity, [86–87](#)

Contingency tables, [137–139](#) , [138t](#)

χ^2 and, [137–139](#) , [138t](#)

Continuous data

effect size, *See* [Effect size](#)

grouped frequency distributions, [96–97](#) , [97t](#)

organization and presentation, [96](#)

Control groups, [43](#)

critical evaluation of published research, [161](#)

definition of, [46](#)

internal validity and, [43](#)

no intervention control group, [46](#)

placebo effect and, [42–43](#)

Control variables, naturalistic non-RCT research designs, [50](#)

Controlled observation, [5](#)

Correlation, [114–120](#) , [115t](#)

- associations between variables, 118–119 , 118f
- causation and, 119–120
- correlation coefficient, 114 , 116–118 , 116t
 - calculation of, 116–118
 - inter-observer (inter-rater) reliability, 119
 - Pearson's r , 116–117 , 117f , 117t
 - predictive validity, 119 , 119t
 - reliability, 119 , 119t
 - selection of, 116
 - shared variance, estimation of, 119
 - Spearman's ρ , 117–118 , 118t
 - test-retest reliability, 119
- in health research, 118–119
- line of 'best-fit', 115
- linear correlation, 115 , 115f
- negative correlation, 115
- nonlinear correlation, 115 , 115f
- positive correlation, 115
- scattergram/scatterplot, 115 , 115f
- Correlational designs, 52–53 , 53t
- 'Counterbalanced' design, 44
- Credibility, 155
- Critical evaluation of published research, 157–163
 - aims or hypotheses, 160
 - discussion, 162–163

- generalization of results, 163
- inferences, 162–163
- interpretations, 162
- research protocol deviations, 162–163
- statistical and clinical significance, 163
- theoretical significance, 163
- guidelines, 158
 - problems, 159t–160t
 - quantitative approaches, 158
 - websites, 158
- instruments/apparatus/tools for, 161
 - validity and reliability of, 161
- literature review, 158–160
- methods section, 161–162
- participants, 161
 - description of study, 161
 - sample size, 161
 - sampling model, 161
- procedure, 161–162
 - assignment, 161–162
 - control groups, 161
 - settings, 162
 - treatment parameters, 162
- research design, 158–160
- research strategy selection, 160

results section, [162](#)

qualitative analysis, [162](#)

statistics, [162](#)

tables and graphs, [162](#)

Critical values, [111](#) , [111f](#)

Cultural differences, qualitative research on, [58](#)

Cumulative frequency, [103](#)

D

Data

analysis, [14](#) , [61](#) , [148](#)

computer-assisted, *See* [Computer-assisted data analysis](#)

critical evaluation, *See* [Critical evaluation of published research](#)

descriptive statistics, *See* [Descriptive statistics](#)

inferential statistics, *See* [Inferential statistics](#)

qualitative, *See* [Qualitative research](#)

quantitative, *See* [Quantitative research](#)

coding of data, *See* [Coding, of data](#)

collection methods

focus groups, *See* [Focus groups](#)

interviews, *See* [Interviews](#)

observation, *See* [Observation](#)

qualitative research, [61](#)

questionnaires, *See* [Questionnaires](#)

retrospective, [52](#)

interpretation of, *See* [Interpretation of data](#)

interviews, [80–82](#)

organization and presentation of, *See* [Organization and presentation of data](#)

quantitative and qualitative methods, [10](#)

saturation, [155](#)

types

continuous, *See* [Continuous data](#)

discontinuous, *See* [Discontinuous data](#)

discrete, *See* [Discrete data](#)

interval data, [96–99](#)

nominal, *See* [Nominal data](#)

ordinal, *See* [Ordinal data](#)

ratio, [96–99](#)

Decision level, α (alpha), [133](#)

Decision making, *See* [Clinical decision making](#)

Declaration of Helsinki (1964, WMA), [35](#)

Degrees of freedom, [129](#)

Demographic questions, [69](#)

Deontological ethics, [34](#)

Dependent variables, [41](#)

multiple, [45](#)

Description, of phenomena, [4](#)

Descriptive epidemiology, [51](#)

Descriptive statistics, [13–14](#) , [102–107](#) , [148](#)

discontinuous data, [99–101](#)

odds, [100–101](#) , [100t–101t](#)

percentages, [99](#)

proportions, [99](#)

rates, [99–100](#)

ratios, [99](#)

inferential and, relationship between, [135](#)

mean, [102–103](#)

median, [103–104](#) , [103f](#) , [103t](#)

mode, [103–104](#) , [103t–104t](#)

normal distribution, *See* [Normal distribution](#)

standard scores, *See* [Standard scores](#)

variability, *See* [Variability](#)

Design, qualitative research and, [61](#)

Determinism, [3](#)

Deviant case sampling, [31](#)

Deviations

average, [105](#) , [105t](#)

standard, [106](#)

Discontinuous data, [99–101](#)

odds, [100–101](#) , [100t–101t](#)

percentages, [99](#)

proportions, [99](#)

rates, [99–100](#)

ratios, [99](#)

Discrete data, [94–95](#)

graphing, [95–96](#) , [96f](#) , [96t](#)

nominal data, *See* [Nominal data](#)

ordinal data, *See* [Ordinal data](#)

organization, [95](#)

Discussion (research publications), [16](#)

critical evaluation, *See* [Critical evaluation of published research](#)

Dopamine (DA)-expressing cells, [181](#) , [182f](#)

Double-barrelled questions, [69](#)

Double-blind design, [47](#)

E

Ecological designs, [52–53](#)

Ecological validity, [32–33](#)

setting of research and, [162](#)

Effect size, [142–144](#) , [142t](#)

absolute risk reduction, [144](#)

clinical decision making and, [147–148](#) , [148f](#) , [148t](#)

Cohen's *d*, [144](#)

continuous data, [142–143](#)

test-retest reliability, [142](#) , [142t](#)

two treatment group, [143](#) , [143t](#)

discontinuous data, [144](#)

odds ratio, [144](#)

randomized controlled trial (RCT), [144](#)

meta-analysis, [172–173](#)

Empathy, [153](#)

Empiricism, [3](#)

Endemic pattern, [52](#)

Epidemic pattern, [52](#)

Epidemiology, [51–52](#)

- analytical, [51–52](#)

- correlational designs, [52–53](#) , [53t](#)

- descriptive, [51](#)

- incidence of, [51](#)

- prevalence of, [51](#)

Equipment, availability of, [37](#)

Equivalence, [89](#)

Errors

- sample size, [33](#)

- sampling, [30–31](#) , [30f](#)

- type I, *See* [Type I error](#)

- type II, *See* [Type II error](#)

Ethical considerations, in health research, [24](#)

- anonymity, [36](#)

- benefits of the research, [38](#)

- clinical research, [66](#)

- confidentiality and right to privacy, [36](#)

- conflict of interest, [36](#)

- independent review, [37](#)

- informed consent, *See* [Informed consent](#)

- key principles and concepts, [35–37](#)
- no intervention control group, [46](#)
- placebo effect, [42](#)
- qualitative research and, [61](#)
- questionnaires, [66–69](#)
- research
 - design and, [34](#)
 - funding bodies, [37](#)
 - proposals, *See* [Research](#)
 - question, [38](#)
- resources, [37–39](#)
- risk and harm minimization, [36](#)
- scientific excellence and quality, [35–36](#)

Ethics, [34–39](#)

- bioethics, [34](#)
- history of modern health research, [35](#)
- medical, [34](#)
- philosophical principles of, [34–37](#)
 - deontology, [34](#)
 - utilitarianism, [34–35](#)

Ethics committees, [38–39](#)

- members, [38](#)

Ethnography, [81](#)

Ethnomethodology, [59](#)

Evidence

interpretation of, [14](#)

quantitative and qualitative methods, [10](#)

Evidence-based health care, [16–17](#)

Evidence-based medicine hierarchy, [50f](#)

Evidence-based practice

 Cochrane Collaboration, [175](#)

 confidence intervals, [130–131](#)

 systematic reviews, [175](#)

Expectancy effects, [46](#) , [161–162](#)

Expected frequency (f_e), [139](#)

Experimental designs

 confounding and bias, *See* [Confounding and bias](#)

 control groups, *See* [Control groups](#)

 dependent variable, [41](#)

 double-blind design, [47](#)

 epidemiology, *See* [Epidemiology](#)

 factorial design, *See* [Factorial design](#)

 independent variable, [41](#)

 post-test-only design, [44](#)

 pre-test/post-test design, [44](#)

 randomized controlled trials and, [40–48](#)

 repeated measures design, [44](#)

 types of, [44–45](#)

Expert opinion, [49](#)

Expertise, availability of, [37](#)

External validity, [26–33](#)
 ecological validity, [32–33](#)
 of experiments and RCTs, [47–48](#)
 of naturalistic designs, [55–56](#)
 population validity, [32](#)
 sampling and, [32–33](#)
Extraneous variables, [52](#)
Extreme response mode, [68](#)

F

Face-to-face interviews, [76–77](#)
Face validity, [86–87](#)
Facilitator, focus groups, [78–79](#)
Factorial design, [45](#) , [45t](#)
 independent groups and, [45](#) , [45f](#)
Factual questions, [69](#)
Falsification, [6](#)
Field notes, [10](#)
Focus groups, [78–80](#) , [80t](#)
 facilitator, [78–79](#)
 participants, [78–79](#)
Follow-up, interviews, [78](#)
Forced-choice response format, [68](#) , [68t](#)
Frequency distributions
 graphing, *See* [Graphing](#)

grouped, [96–97](#) , [97t](#)

Frequency polygons, [98–99](#) , [98f](#)

Friedman two-way analysis of variance, [136t](#)

Fully crossed design, [45](#)

Funding bodies, research, [36–37](#)

G

Gaussian curve, [109](#)

Generalization, critical evaluation of published research, [163](#)

Graphing, [148](#)

bar graph, [95](#) , [96f](#)

discrete data, [96f](#) , [96t](#)

frequency distributions, [98–99](#)

frequency polygons, [98–99](#) , [98f](#)

histograms, [98](#) , [98f](#)

pie chart, [96](#) , [96f](#) , [96t](#)

Grounded theory, [81–82](#) , [153–154](#)

example, [154](#)

See also [Thematic analysis](#)

Group interview, *See* [Focus groups](#)

Grouped frequency distributions, [96–97](#) , [97t](#)

worked example, [97](#)

Groups, assignment, *See* [Assignment](#)

H

Harm, *See* [Risk and harm minimization](#)

Health care

- evidence-based, [16–17](#)

- qualitative evidence, application, [11](#) , [57](#)

- quantitative evidence, application, [11](#)

Health care environments, evaluation of, [58](#)

Health problems, example, [71–73](#)

Health research

- clinical decision making, *See* [Clinical decision making](#)

- confidence intervals, use, [130–131](#)

- correlations, use in, [118–119](#)

- evidence-based health care and, [16–17](#)

- foundations of, [1–7](#)

 - knowledge, [3](#)

 - method, [3](#)

- qualitative topics, [60](#)

- research problems, *See* [Research](#)

- scientific methods, *See* [Scientific methods](#)

- social context, *See* [Social context](#)

Hermeneutics, [154](#)

Histograms, [98](#) , [98f](#)

History, [42](#)

Holistic perspective, [10](#)

Hypotheses, [24](#) , [126](#) , [132](#)

- critical evaluation of published research, [160](#)

Hypothesis testing, 132–140

‘alternative hypothesis’, 133

decision level, α (alpha), 133

illustration of, 133–134 , 134f

null hypothesis, 133

probability, 133

sampling distribution, 134

steps, 133

types I error, *See* Type I error

types II error, *See* Type II error

I

Immersion, 156

Incidence rates, 51 , 99 , 100f

Incidental samples

incidental sampling, 27–28

quota sampling, 28 , 28t

Inclusion criteria, 29 , 168 , 174

Independent variables, 41

design, 45 , 45f

levels of, 44t

Inferential statistics, 14 , 123 , 148

chi-square test, *See* Chi-square test

confidence intervals, *See* Confidence intervals

degrees of freedom, 129

descriptive statistics and, relationship between, [135](#)

hypothesis testing, *See* [Hypothesis testing](#)

nonparametric tests, [136](#)

not required, [135](#)

parametric tests, [136](#)

probability, *See* [Probability](#)

statistical packages, *See* [Statistics](#)

test, selection, examples, [136](#)

type I error, *See* [Type I error](#)

type II error, *See* [Type II error](#)

Informed consent, [35](#)

plain language statements, [35](#)

Inter-observer (inter-rater) reliability, [86](#) , [86t](#) , [119](#)

Internal consistency, [86](#)

Internal validity, [43](#) , [48](#)

control groups and, [43](#)

multiple-group time-series designs, [55](#)

natural recovery and, [43](#)

of naturalistic designs, [55–56](#)

Interpolation, [98](#)

Interpretation of data

critical evaluation of published research, [162](#)

meta-analysis, *See* [Meta-analysis](#)

qualitative data, *See* [Qualitative research](#)

systematic reviews, [171](#)

type I error, *See* [Type I error](#)

type II error, *See* [Type II error](#)

using statistical packages, [46](#)

Interquartile range, [106](#)

Interval data, organization and presentation, [96–99](#)

Interval scales, [89–90](#) , [90t](#)

Interventions, evaluation of, [58](#)

Interviews, [75](#)

analysis of data, [80–82](#)

coding, [82](#)

ethnography, [81](#)

grounded theory, [81–82](#)

qualitative, [81–82](#)

quantitative, [81](#)

thematic analysis, [82](#)

focus groups, *See* [Focus groups](#)

methods, [76–77](#)

face-to-face, [76–77](#)

remote, [76–77](#)

models, [75–76](#) , [76t](#)

non-schedule standardized interview, [75](#)

schedule standardized interview, [75](#)

structure/formality, [75](#)

note-taking, [78](#)

process, [77–78](#)

- follow-up, [78](#)
- free-form (unstructured) notes, [77–78](#)
- participants, [77](#)
- response schedules/answer checklists, [77](#)
- questionnaires, compared with, [66](#) , [66t](#)
- recording information, [78–80](#)
 - advantages and disadvantages of, [78](#) , [79t](#)
 - audiotaping, [78](#)
 - video recording, [78](#)
- techniques, [75–82](#)
- transcript, [80](#)
- Introduction (research publications), [15](#)
 - See also* [Literature review](#)
 - critical evaluation, *See* [Critical evaluation of published research](#)

K

Kruskal-Wallis H test, [136t](#)

L

Likert scales, [68](#) , [68t](#)

Line of ‘best-fit’, [115](#)

Linear correlation, [115](#) , [115f](#)

Literature review, [22–23](#)

- critical evaluation of published research, [158–160](#)
- literature searching, [22–23](#)

PICO approach, [22](#)
synthesis of research, [166](#)
“Lost in translation”, [181–183](#)

M

Mammography, case study of test validity, [87](#)
Mann-Whitney *U* test, [136t](#)
Matched group, [44](#)
Maximum variation sampling, [31](#)
McNemar’s test, [136t](#)
Mean, [102–103](#)
Measurement, [83–91](#)

- definition, [84](#)
- operationalization, [84](#) , [85f](#)
- reliability, *See* [Reliability](#)
- standardized measures and tests, *See* [Standardized measures and tests](#)
- tools and procedures, [85–88](#)

Measurement scale types, [89–90](#) , [90t](#)

- interval scales, [89–90](#) , [90t](#)
 - nominal scales, [89](#) , [89t–90t](#)
 - ordinal scales, [89](#) , [89t–90t](#)
- ratio scales, [90](#) , [90t](#)

Median, [103–104](#) , [103f](#) , [103t](#)
Medical ethics, [34](#)
Meta-analysis, [172–175](#)

combining data from diverse studies, [172–173](#) , [172t](#)

extracting data from individual studies, [172](#)

effect size, [172](#)

forest plot, [173](#)

interpretation of results, [173–174](#) , [173f–174f](#)

confidence intervals, [173](#)

effect size, [173](#)

odds ratio, [174](#)

sample sizes, [174](#)

studies, [173](#)

weights and totals, [174](#)

strategies, [172](#)

qualitative, [172](#)

quantitative, [172](#)

software packages, [172](#)

Method(s) (health research), [2–3](#)

scientific, [3–4](#)

Method(s) section (research publications), [15–16](#)

apparatus and tools, [15](#)

critical evaluation, *See* [Critical evaluation of published research](#)

participants, [15](#)

procedure, [15–16](#)

Mixed designs, [71](#) , [155](#) , [160](#)

Mixed methods, [71–74](#)

examples of, [72–73](#) , [72t](#)

qualitative research, [72–73](#) , [73f](#)

quantitative research, [72–73](#) , [73f](#)

Mode, [103–104](#) , [103t–104t](#)

Modern health research ethics

Declaration of Helsinki (1964, WMA), [35](#)

history of, [35](#)

Nuremberg Code (1947), [35](#)

Monomethod approaches, [71](#)

Moral conduct, [34](#)

Moral principles, [34](#)

Multiple dependent variables, [45](#)

Multiple-group time-series design, [54–55](#) , [55f](#)

Murphy's Law, [38](#)

N

Narrative reviews, [167](#)

quality, [167](#)

Natural comparison studies, [52](#)

Naturalistic non-RCT research designs, [50–51](#)

Negative correlation, [115](#)

Neural plasticity, [183](#)

Neural reconstruction, [181](#)

Neurologic problems, relating to people with, [58](#)

Nominal data, organization and presentation, [94–96](#) , [95t](#)

Nominal scales, [89](#) , [89t–90t](#)

Non-schedule standardized interview, [75](#)
Non-standardized interview, [75](#)
‘Nondirectional’ probability, [130](#)
Nonequivalence, [89](#)
Nonlinear correlation, [115](#) , [115f](#)
Nonparametric tests, [136](#)
Normal curve, *See* [Normal distribution](#)
Normal distribution, [108–113](#) , [109f](#)
 calculating area under normal curve, [110–111](#)
 critical values, [111](#) , [111f](#)
 standard normal curve, [110](#)
 for comparison of distributions, [111–112](#) , [112f](#) , [112t](#)
 worked example, [110](#)
Note-taking, [78](#)
Null hypothesis, [133](#) , [145](#)
 rejecting, [133](#)
Null results, [146](#) , [146t](#)
Nuremberg Code (1947), [35](#)

O

Observation, [83–91](#)
Observer(s)
 bias, [42](#)
 inter-observer (inter-rater) reliability, [86](#) , [86t](#)
Odds ratio (OR), [100t–101t](#) , [101](#) , [144](#)

- meta-analysis interpretation of results, [174](#)
- Open-ended question formats, [67–68](#) , [67t](#)
- Operational definition, [84](#)
- Operationalization, operational definition, [84](#) , [85f](#)
- Operationalize constructs, [84–85](#) , [85f](#)
- Opinion questions, [69](#)
- Ordinal data, organization and presentation, [94–96](#) , [95t](#)
- Ordinal scales, [89](#) , [89t–90t](#)
- Organization and presentation of data, [13–14](#) , [92–101](#)
 - continuous data, *See* [Continuous data](#)
 - discrete data, *See* [Discrete data](#)
 - interval data, [96–99](#)
 - nominal data, *See* [Nominal data](#)
 - ordinal data, *See* [Ordinal data](#)
 - ratio data, [96–99](#)
- Outcome prediction studies, [52–53](#)
- Outcomes, measurement, [88](#)

P

- Pandemic pattern, [52](#)
- Paradigm, [4](#)
 - positivist, [4–5](#)
 - controlled observation, [5](#)
 - deduction, [5](#)
 - generalization and induction, [4](#)

- health care, research and advances of, [5](#)
- hypotheses and laws, [4](#)
- observations, description and measurement, [4](#)
- theories, [4–5](#)
- verification and falsification, [5](#)

- post-positivist, [5–7](#)
 - falsification, constitutes, [6](#)
 - social context, science in, [6–7](#)
 - theory-dependence of observation, [6](#)

Parametric tests, [136](#)

Parkinson's disease (PD), [71](#) , [181](#)

Participants, [15](#)

- assignment into groups, *See* [Assignment](#)
- availability of, [37](#)
- critical evaluation of published research, *See* [Critical evaluation of published research](#)
- ethical considerations, *See* [Ethical considerations](#) , [in health research](#)
- focus groups, [78–79](#)
- human, [47](#)
- interview, [77](#)
- participant observer, [38](#)
- placebo effect, *See* [Placebo effect](#)
- population definition, [45](#)
- randomized-controlled trial (RCT) in, [45–47](#)
- rights and benefits from research, [38](#)
- sampling, [23](#)

standardized measures and tests, 88

Path analysis, 120

Pearson's r , 116–117 , 117f , 117t

assumptions, 117

Percentages, 99

Percentiles, 106 , 107f , 111

Personal meaning, 60

Phase 0 Clinical Trials, 180

Phase I Clinical Trials, 180

Phase II Clinical Trials, 180–181

Phase III Clinical Trials, 181

Phase IV Clinical Trials, 181

Phenomenology, 59

PICO approach to literature reviews, 22

Pie chart, 96 , 96f , 96t

Pilot study, 24

questionnaires, 66

Placebo effect, 43

control groups and, 43

double-blind design, 47

Population

definition of, 45

health indicators of, 51

validity, 32

Positioning, of researcher, 10

Positive correlation, [115](#)

Positivist paradigm, [4–5](#)

- controlled observation, [5](#)

- deduction, [5](#)

- generalization and induction, [4](#)

- health care, research and advances of, [5](#)

- hypotheses and laws, [4](#)

- observations, description and measurement, [4](#)

- theories, [4–5](#)

- verification and falsification, [5](#)

Post-positivist paradigm, [5–7](#)

- falsification, constitutes, [6](#)

- social context, science in, [6–7](#)

- theory-dependence of observation, [6](#)

Post-test only design, [44](#)

Power, analysis, [146–147](#) , [147t](#)

Pragmatism, [7](#)

Pre-test/post-test design, [44](#)

Preclinical research, [180–181](#)

Predictive validity, [87](#) , [87t](#) , [119](#) , [119t](#)

Prevalence rates, [51](#) , [99](#)

PRISMA flow chart review, [168f](#)

Probability, [123–125](#) , [124f–125f](#) , [124t](#)

- hypothesis testing, *See* [Hypothesis testing](#)

- nondirectional, [130](#)

sampling distribution and, [125–126](#) , [125f–126f](#)

sampling distributions, *See* [Sampling](#)

two-tail, [130](#)

Procedure (research publications), critical evaluation, *See* [Critical evaluation of published research](#)

Professional publication, [14–15](#)

Proportions, [99](#)

Protocol deviations, critical evaluation of published research, [162–163](#)

Psychiatric problems, relating to people with, [58](#)

Publications, *See* [Research publications](#)

Purposive sampling, [31–32](#) , [32t](#)

strategies, [31](#)

extreme or deviant case sampling, [31](#)

maximum variation sampling, [31](#)

snowball or chain sampling, [31–32](#) , [32t](#)

theory-based sampling, [32](#)

Q

Qualitative methods, [8–11](#) , [23](#) , [60](#)

quantitative methods vs., [9–11](#) , [9t](#)

data and evidence, [10](#)

health care delivery, applications in, [11](#)

research, positioning of, [10](#)

subject matter, perception of, [9–10](#)

theories, [10–11](#)

theory testing, [11](#)

reality, social construction of, 8–9

Qualitative research, 7 , 73–74

analysis of data, 81–82 , 150–156

accuracy, 155 , 155f

coding, *See* Coding, of data

computer-assisted data analysis, 153

content analysis, *See* Content analysis

critical evaluation, 162

thematic analysis, *See* Thematic analysis

worked example, 154

applications in health care, 57

approaches to, 59–60

ethnomethodology, 59

grounded theory, 59–60

phenomenology, 59

symbolic interactionism, 59

auditability, 155

credibility, 155

on cultural differences, 58

data saturation, 155

define, 60–61

ethnography, 81

example, 61–62

grounded theory, *See* Grounded theory

immersion, 156

interpretation of data, [151](#) , [154–155](#)

semiotics, [154](#)

worked example, [154](#)

interviews, *See* [Interviews](#)

meta-analysis strategy, [172](#)

mixed research studies, [72–73](#) , [73f](#)

observation, *See* [Observation](#)

parallel, [73–74](#)

planning, [61](#)

problems and questions, [58–59](#)

questionnaires, *See* [Questionnaires](#)

sampling in, [31](#)

social context, [154–155](#)

thematic analysis, *See* [Thematic analysis](#)

triangulation, [155](#) , [155f](#)

Quantitative methods, [8–11](#) , [23](#)

qualitative methods vs., [9–11](#) , [9t](#)

data and evidence, [10](#)

health care delivery, applications in, [11](#)

research, positioning of, [10](#)

subject matter, perception of, [9–10](#)

theories, [10–11](#)

theory testing, [11](#)

reality, social construction of, [8–9](#)

reductionism, [9–10](#)

Quantitative research, [7](#) , [73–74](#) , [81](#)

inferential statistics, *See* [Inferential statistics](#)

interviews, [76](#) , [81](#)

meta-analysis strategy, [172](#)

mixed research studies, [72–73](#) , [73f](#)

observation, *See* [Observation](#)

parallel, [73–74](#)

questionnaires, *See* [Questionnaires](#)

statistics, *See* [Statistics](#)

Quartile deviation, [106–107](#)

Quasi-experimental designs, [49–56](#)

multiple-group time-series designs, [54–55](#) , [55f](#)

time-series cohort designs, [54](#) , [54f](#)

Questionnaires, [63–70](#) , [68t](#)

administration

interviewers, [66](#)

administration of, [65](#)

clinical research, use of in, [66](#)

construction, [64–66](#)

drafting of, [65](#)

interviews, compared with, [66t](#)

Likert scales, [68](#) , [68t](#)

pilot study, [66](#)

piloting/testing of, [65](#)

question format, [66–69](#) , [66t](#)

- ambiguous questions, 69
- biased and leading questions, 69
- closed-response question, 67–68 , 67t
- double-barrelled questions, 69
- forced-choice response, 68 , 68t
- level of wording, 69
- open-ended questions, 67–68 , 67t
- redrafting/finalization of, 65
- structure of, 69
 - closing statements and return instructions, 69
 - demographic questions, 69
 - factual questions, 69
 - introductory statement, 69
 - opinion questions, 69

Quota sampling, 28 , 28t

R

Random assignment, 46

Random sampling, 28–29

- advantages of, 29

- disadvantages of, 29

- list method, 29

- procedure of, 28

- raffle method, 28–29

- stratified random sampling, *See* Stratified random sampling

Randomized-controlled trial (RCT), [45–47](#) , [45f](#)

- assignment, [46](#)

- control groups, [46](#)

- defining, [41](#)

- double-blind design, [47](#)

- efficacy, [46](#)

- external validity of, [47–48](#)

- outcome measurement, [46](#)

- statistical analysis and interpretation, [46](#)

Range, [105](#)

Rank-ordering, of scores, [118t](#)

Rates, [99–100](#)

- incidence rates, [99](#) , [100f](#)

- prevalence rates, [99](#)

Ratio data, organization and presentation, [96–99](#)

Ratio scales, [90](#) , [90t](#)

Ratios, [99](#)

Raw scores, [109](#) , [111](#)

Realism, [3](#)

Reasoning, [3](#)

Reconceptualized model, of translational research, [178](#) , [178f](#)

Recording information, [78–80](#) , [79t](#)

Reductionism, [9–10](#)

References (research publications), [16](#)

Reliability, [86](#)

correlation, [119](#) , [119t](#)

critical evaluation of published research, [161](#)

inter-observer (inter-rater) reliability, [86](#) , [86t](#) , [119](#)

internal consistency, [86](#)

standardized measures and tests, *See* [Standardized measures and tests](#)

test-retest reliability, [86](#) , [119](#) , [142](#) , [142t](#)

Remote interviews, [76–77](#)

Repeated measures design, [44](#)

Representative samples, [27–29](#) , [27f](#)

Research

 aims, [24](#)

 critical evaluation of published research, [160](#)

 basic, [177](#)

 benefits of, [38](#)

 clinical, [177](#)

 phases of, [180–181](#)

 collection methods, [13](#)

 design, [13](#)

 ‘before-after’ design, [42](#) , [42f](#)

 critical evaluation of published research, [161](#)

 experimental designs, *See* [Experimental designs](#)

 funding bodies, ethics and, [36–37](#)

 hypotheses, [24](#)

See also [Hypotheses](#)

 planning, [13](#)

- economic issues, [37](#)
- qualitative research, [61](#)
- problems, [13](#) , [20–22](#) , [21t](#)
 - cost-effectiveness of health services, [22](#)
 - demographic changes, [21–22](#)
 - environmental and social changes, [20–21](#)
 - qualitative research, an example, [61](#)
 - scientific and technological advances, [22](#)
- process, [12–17](#)
 - data analysis, *See* [Data](#)
 - data collection, *See* [Data](#)
 - data organization, *See* [Organization and presentation of data](#)
 - data presentation, *See* [Organization and presentation of data](#)
 - dissemination, [14](#)
 - evaluation and dissemination, [14](#)
 - literature review, *See* [Literature review](#)
 - sequential steps of, [12](#) , [13f](#)
- proposals, [24](#)
- qualitative method in, [23](#)
- quantitative method in, [23](#)
- questions, [13](#) , [60](#)
 - ethical considerations and, [38](#)
 - formulation of, [18–25](#) , [21f](#)
 - methodological considerations, [23–24](#)
 - qualitative research, an example, [61](#)

- systematic reviews, [169](#)
- strategy selection, [160](#)
- study design, [23–24](#)
- translational, [177–184](#)
 - define, [177–178](#)
 - example of obstacles to successful, [181–183](#)
 - reasons for, [178–180](#) , [178f](#)
 - stages of, [179–180](#) , [179f](#)

Research publications, [14–16](#) , [14t](#)

- critical evaluation, *See* [Critical evaluation of published research](#)
- discussion, [16](#)
- introduction, [15](#)
- meta-analysis, *See* [Meta-analysis](#)
- method, [15–16](#)
- publication process, [16](#)
- references and appendices, [16](#)
- results section, [16](#)
- style, [17](#)
- systematic reviews, *See* [Systematic reviews](#)
- title and abstract, [15](#)

Resources available, [37](#)

Results (research publications), [16](#)

- analysis of data, *See* [Data](#)
- critical evaluation of results section, *See* [Critical evaluation of published research](#)
- organization and presentation, *See* [Organization and presentation of data](#)

Return instructions, [69](#)

Risk and harm minimization, [36](#)

S

Sample

- selection of, [45](#)

- size, [30–31](#) , [125](#)

 - critical evaluation of published research, [161](#)

 - meta-analysis interpretation of results, [174](#)

Sampling

- area sampling, [29](#)

- critical evaluation of published research, [161](#)

- distribution, [125–126](#) , [125f–126f](#)

 - central limit theorem, [126](#)

 - of mean, [126–127](#) , [127f](#)

- error, [30–31](#) , [30f](#)

- methods, [26–33](#)

 - aim, [33](#)

 - basic issues in, [26–27](#)

 - incidental sampling, *See* [Incidental samples](#)

 - populations and, [28t](#)

 - purposive, *See* [Purposive sampling](#)

 - in qualitative research, *See* [Qualitative research](#)

 - random sampling, *See* [Random sampling](#)

 - representative sample, [27–29](#) , [27f](#)

- stratified random sampling, *See* [Stratified random sampling](#)
- types, [33](#)
- systematic reviews, [174](#)
- systematic sampling, [29](#)
- Scattergram (scatterplot), [115](#) , [115f](#)
- Scepticism, [3](#)
- Schedule standardized interview, [75](#)
- Scientific methods
 - history of, [3–4](#)
 - positivist paradigm, [4](#) , [4f](#)
- Scientific Revolution, [3](#)
- Search strategy, systematic reviews, [170](#)
- Semi-interquartile range, [106–107](#) , [107f](#)
- Semiotics, [154](#)
- Sensitivity, [87](#) , [87t](#)
- Shared variance, estimation of, [119](#)
- Sign test, [136](#) , [136t](#)
- Significance, [141](#)
 - confidence intervals and, [144–145](#) , [145f](#)
 - critical evaluation of published research, [163](#)
- Snowball sampling, [31–32](#) , [32t](#)
- Social context, of research
 - qualitative research, [154–155](#)
 - science in, [6–7](#)
- Spearman's ρ , [117–118](#) , [118t](#)

Specificity, [87](#) , [87t](#)

SPSS, *See* [Statistical Package for Social Sciences](#)

Standard deviation, [106](#)

Standard scores (Z scores), [108–113](#) , [109f](#)

Standardized measures and tests, [88](#)

- assessment tool, [88](#) , [89b](#)

- acceptability to participants/patients, [88](#)

- accessibility, [88](#)

- reliability, [88](#)

- theoretical framework, [88](#)

- validity, [88](#)

Statistical inferences, optimize, [145–146](#) , [145f](#)

Statistical Package for Social Sciences (SPSS), [139](#)

Statistical significance, [141](#)

Statistics, [94](#)

- descriptive, *See* [Descriptive statistics](#)

- inferential, *See* [Inferential statistics](#)

- statistical packages, [139–140](#)

- descriptive statistics, [139t](#) , [140](#)

- encoding raw data, [139](#)

- identifying the statistical analysis required, [139](#)

- inferential statistics, [139](#)

- interpretation of, [46](#) , [139](#)

- printout, [139](#) , [139t–140t](#)

- selection, [139](#)

training, [139](#)

Stratified random sampling, [29](#)

advantages of, [29](#)

disadvantages of, [29](#)

Style guide for research publications, [17](#)

Survey design, [63–70](#)

Surveys, [51](#)

Symbolic interactionism, [59](#)

Synthesis of research, [164–176](#)

basic principles, [166](#)

literature review, [166](#)

meta-analysis, *See* [Meta-analysis](#)

narrative reviews of research, *See* [Narrative reviews](#)

systematic review, *See* [Systematic reviews](#)

Systematic reviews, [49](#) , [167–171](#)

appraisal and selection of key papers, [170](#)

criteria example, [170](#)

characteristics, [167–171](#)

Cochrane Collaboration, *See* [Cochrane Collaboration](#)

coding, [170–171](#)

features, [170](#) , [171t](#)

evidence-based practice, [175](#)

interpretation of research evidence, [171](#)

publication of, [171](#)

research questions, [169](#)

sampling, 174

See also [Synthesis of research](#)

search strategy, 170

steps in, 167–169

- assessing quality of, 168 , 168f

- framing questions, 167

- identifying relevant work, 168

- interpreting findings, 169

- summarizing evidence, 168 , 169f

theoretical framework, 169

validity, 174–175

- extracting the data, 174–175

Systematic sampling, 29

T

t distribution, 128–130 , 129f–130f

Tables of results, critical evaluation, 162

Tabulating data, conventions, 95

Telephone interviews, 76–77

Test validity, types, 86–87

- construct validity, 87

- content or face validity, 86–87

Test-retest reliability, 86 , 119 , 142 , 142t

Thematic analysis, 82 , 152

- example, 154

themes, [154](#)

See also [Grounded theory](#)

Theoretical framework, systematic reviews, [169](#)

Theories

qualitative methods, [10–11](#)

quantitative methods, [10–11](#)

Theory-dependence of observation, [6](#)

Theory testing, [11](#)

Time, availability of, [38](#)

Time-series cohort designs, [54](#) , [54f](#)

Title (research publications), [15](#)

Tradition, [3](#)

Traditional Chinese Medicine (TCM), [3](#)

Transcript, [80](#)

Translation ‘pipeline’, [179](#) , [179f](#)

Translational Medicine, [179](#)

Translational research, [177–184](#)

basic research, [177](#)

clinical research, [177](#)

phases of, [180–181](#)

define, [177–178](#)

example of obstacles to successful, [181–183](#)

reasons for, [178–180](#) , [178f](#)

stages of, [179–180](#) , [179f](#)

discovery to candidate health application, [179–180](#)

evidence-based guidelines to health practices, [180](#)

health applications to evidence-based guidelines, [180](#)

practice to population health impact, [180](#)

Treatment

administration of intervention, [46](#)

fidelity, [162](#)

parameters, [162](#)

Triangulation, [74](#) , [155](#) , [155f](#)

‘Two-tail’ probability, [130](#)

Type I error, [133–134](#) , [145](#)

clinical decision making, [145f](#) , [147–148](#) , [148t](#)

occurrence, minimize, [145](#)

type II error and, [145](#)

Type II error, [133–134](#) , [145](#)

clinical decision making, [147–148](#) , [148f](#) , [148t](#)

type I error and, [145](#)

U

Unstructured interview, [75–76](#)

Utilitarianism, [34–35](#)

V

Validity, [86–88](#)

case study of test validity, [87–88](#)

confounding and bias, *See* [Confounding and bias](#)

critical evaluation of published research, [161](#)

ecological, *See* [Ecological validity](#)

external, *See* [External validity](#)

internal, *See* [Internal validity](#)

predictive validity, [87](#) , [87t](#) , [119](#) , [119t](#)

systematic reviews, *See* [Systematic reviews](#)

test validity, types, [86–87](#)

 construct validity, [87](#)

 content or face validity, [86–87](#)

 predictive validity, [87](#) , [87t](#)

 sensitivity and specificity, [87](#) , [87t](#)

Variability, [102–107](#)

 percentiles, [106](#) , [107f](#)

 range, [105](#)

 average deviation, [105](#) , [105t](#)

 semi-interquartile range, [106–107](#) , [107f](#)

 standard deviation, [106](#)

See also [Variance](#)

Variables

 control variables, natural comparison studies, [52](#)

 dependent, *See* [Dependent variables](#)

 independent, *See* [Independent variables](#)

Variance, [105–106](#) , [106t](#)

 shared variance, estimation, [119](#)

Verstehen, [153–154](#)

Video recording, [78](#)

Volunteer effect, [29](#)

W

Weight of study, [174](#)

Wording of questions, [69](#)

Z

Z scores (standard scores), [108–113](#) , [109f](#)