CHAPTER 1

Frameworks for Generating and Applying Evidence

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the purpose of clinical research.
2. Describe how quantitative and qualitative research differ and how they complement each other.
3. Describe the steps of the research process.
4. Discuss the role of evidence in clinical decision making.
5. Discuss the components of the *International Classification of Functioning, Disability and Health* and how they influence research questions.
6. Explain the role of interprofessional research.
7. Describe the purpose of explanatory, exploratory, and descriptive research.

■ Key Terms

Qualitative research
Quantitative research
Scientific method
Translational research
Efficacy
Effectiveness
Translational research
Efficacy
Effectiveness
International Classification of Functioning, Disability and Health (ICF)
Body structures
Body functions
Activities
Participation
Environmental factors
Personal factors
Intraprofessional
Multiprofessional
Interprofessional
Transprofessional
Basic research
Applied research
Systematic review
Meta-analysis
Scoping review
Explanatory research
Experimental designs
Randomized controlled trial (RCT)
Pragmatic clinical trial (PCT)
Quasi-experimental designs
Single subject designs (SSD)
N-of-1 trial
Exploratory research
Epidemiology
Cohort studies
Case-control studies
Correlational/predictive studies
Methodological research
Descriptive research

Developmental research
Normative research
Case report/case series
Historical research
Mixed methods research

■ Chapter Summary

- Clinical research is a structured process of investigating facts and theories and exploring connections, with the purpose of improving individual and public health.
- NIH defines three clinical research purposes:
  - To conduct patient-oriented research to understand mechanisms of disease and interventions.
  - To conduct epidemiologic observational studies to describe patterns of disease and identify risk factors.
  - To conduct outcomes research and health services research to determine the impact of research on population health and evidence-based interventions.
- Qualitative research strives to capture naturally occurring phenomena, following a tradition of *social constructivism.* This philosophy is focused on the belief that all reality is fundamentally social, and therefore the only way to understand it is through an individual's experience.
- Quantitative research is based on a philosophy of *logical positivism*, in which human experience is assumed to be based on logical and controlled relationships among defined variables. It involves measurement of outcomes using numerical data under standardized conditions.
- The scientific method has been defined as a *systematic, empirical, and controlled critical examination of hypothetical propositions about the associations among natural phenomena.*
- The process of clinical research involves sequential steps that guide thinking, planning and analysis, beginning with identification of the research question, moving to designing and implementing the study, analyzing data, and disseminating findings.
- Evidence-based practice (EBP) involves critical evaluation of literature and consideration of clinical expertise, patient values, and clinical circumstances to inform clinical decision-making.
- Translational research is the application of basic scientific findings to clinically relevant issues, and simultaneously, the generation of scientific questions based on clinical dilemmas.
- Efficacy is the benefit of an intervention as compared to a control, placebo or standard program, tested in a carefully controlled environment, with the intent of establishing cause and effect relationships. Effectiveness refers to the benefits and use of procedures under "real world" conditions where circumstances cannot be controlled within an experimental setting.
- The International Classification of Functioning, Disability and Health (ICF) is a model that shows the relationship among body structures and functions, activities, and participation in life roles. The effect of a health condition on these behaviors can be mitigated by environmental and personal factors.
- Interprofessional research is an important focus to assure that questions are answered

using varied perspectives and expertise.

- Basic research is directed toward the acquisition of new knowledge. Applied research advances the development of new diagnostic tests, drugs, therapies and prevention strategies, answering questions with direct clinical application.
- Explanatory research uses experimental designs to compare two or more interventions. These include randomized controlled trials (RCTs), pragmatic clinical trials (PCTs), and single-subject designs (SSDs).
- Exploratory research uses observational designs to explore relationships and predict risk factors. These include cohort studies, case-control studies and predictive studies for prognosis and diagnostic accuracy.
- Methodological research is used to study reliability and validity of measurements.
- Descriptive research is used to describe groups or populations to document their characteristics. Approaches include developmental and normative research, case reports, historical research, and qualitative research.

CHAPTER 2

On the Road to Translational Research

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Define translational research.
2. Discuss the difference between efficacy and effectiveness.
3. Distinguish characteristics of randomized and pragmatic trials.
4. Describe the translational continuum in research in relation to Phase I-IV trials
5. Discuss the advantages and limitations of comparative effectiveness research in providing evidence for practice.
6. Discuss the relevance of considering patient-reported outcome measures and patient-oriented evidence that matters.
7. Discuss the purpose of implementation studies.

■ Key Terms

Transitional research
Randomized controlled trial (RCT)
Effectiveness trials
Pragmatic clinical trial (PCT)
Basic research
Phase I trials
Phase II trials
Phase III trials
Phase IV trials
Practice based evidence
Outcomes research
Patient-oriented evidence that matters (POEM)
Disease oriented evidence (DOE)
Patient-reported outcome measure (PROM)
Patient-centered outcomes research (PCOR)
Primary outcome
Secondary outcome
Implementation science

■ Chapter Summary

- Translational research refers to the direct application of scientific discoveries into clinical practice, often referred to as taking knowledge from “bench to bedside.”

- Many scientific breakthroughs can take up to 25 years to reach publication and implementation, creating a "translation gap."
- Efficacy is the benefit of a new therapy established using a randomized controlled trial (RCT) which incorporates random allocation, blinding, and designs to reduce sources of bias.
- Effectiveness trials look at the effect of interventions under real world conditions using pragmatic trials.
- Four translation blocks include Phase I trials to determine if treatments are safe in humans, Phase II and III trials to determine efficacy with specific controls and defined patient samples, and Phase IV trials that look at implementation in community settings.
- Comparative effectiveness research (CER) is designed to study real world application of research evidence by comparing two treatments.
- Pragmatic clinical trials (PCTs) study interventions in diverse settings to better reflect practical circumstances of patient care.
- Practice-based evidence (PBE) refers to the type of evidence that is derived from real patient care problems, identifying gaps between recommended and actual practice.
- Outcomes research is an umbrella term that describes the impact of health care practices and interventions.
- Patient-reported outcome measures (PROM) are outcomes that reflect what is most important to patients, such as quality of life and function.
- Patient-centered outcomes research (PCOR) has the goal of engaging patients and other stakeholders in the development of research questions and outcome measures.
- Studies can designate a primary outcome as the one that will be used to assess therapeutic benefit, as well as secondary outcomes that are other endpoint measures.
- Implementation science is the study of methods to promote the integration of research finding and evidence into healthcare policy and practice.

CHAPTER 3

Defining the Research Question

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the process of developing a research question.
2. Discuss the sources of research questions.
3. Describe how a theoretical rationale forms the framework for a research question.
4. Define independent and dependent variables.
5. Describe the purpose of operational definitions.
6. Describe the characteristics of good research hypotheses.
7. Express the purpose of research in terms of the research problem, the research question, specific aims, and research hypotheses.

■ Key Terms

Systematic reviews
Meta-analysis
Explanatory research
Exploratory research
Descriptive research
Methodological research
PICO
Variable
Factor
Independent variable
Dependent variable
Levels
Operational definition
Research hypothesis
Statistical hypothesis
Nondirectional hypothesis
Directional hypothesis
Simple hypothesis
Complex hypothesis
Guiding questions
Specific aims
Power

Chapter Summary

- The research process begins by identifying a topic of interest and clarifying the problem to generate a specific, testable question.
- Questions may arise from clinical experience, from clinical theory, finding gaps or conflicts in the literature, or the need to better understand patterns or characteristics of populations.
- Research questions must be built on a rationale that justifies the need for the study and the foundation for testing it. This process includes a review of literature to understand prior research and the state of knowledge. Systematic reviews and meta-analyses are useful for summarizing the conclusions and quality of past research.
- Research questions can focus on four types of objectives: explaining cause-and-effect through comparison of interventions, looking for relationships to determine how clinical phenomena interact, describing existing conditions or phenomena in a particular population, or studying measurement methods to investigate reliability and validity.
- Research questions should be framed around four essential components using the acronym PICO: the Population being studied, the Intervention of interest (which may be a diagnostic test or risk factor), a Comparison group (if relevant), and Outcomes.
- Studies incorporate independent variables that represent conditions or interventions that predict outcomes, and dependent variables that represent responses or measured outcomes.
- Operational definitions define variables according to their unique meaning within a study design, providing detail of how treatments are applied or outcomes are measured.
- Research hypotheses are declarative statements that predict the relationship between independent and dependent variables, and are used to guide the interpretation of outcomes. The purpose of a study is to test the hypothesis and, ultimately, to provide evidence so that the researcher can accept or reject it. Hypotheses can be expressed with direction (which treatment will be better) or without direction (indicating they will be different).
- Descriptive studies will not have hypotheses, but will use guiding questions or specific aims to frame the study's purpose.
- A good research question will focus on important problems that have an impact on patient care. It will direct a study that adheres to ethical standards and that is feasible in terms of time, expertise, and resources.

CHAPTER

4

The Role of Theory in Research and Practice

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the role of theory in clinical practice and research.
2. Identify four purposes of theories in clinical research.
3. Illustrate deductive and inductive logic and differentiate inductive and deductive theories.
4. Define and illustrate concepts and constructs.
5. Describe propositions and models in relation to clinical theory.
6. Illustrate how hypotheses are derived from theories and are used to test theories.

■ Key Terms

Theory
Concepts
Variables
Constructs
Propositions
Models
Deductive reasoning
Inductive reasoning
Inductive theories
Middle-range theory
Grand theory
Meta-theory
Law

■ Chapter Summary

- A theory is a set of interrelated concepts that specify relationships among variables, representing a reasonable explanation of the relationships.
- Theories are used to summarize knowledge to explain observable events, to predict what should occur under specific conditions, to stimulate development of new knowledge, and to provide a basis for asking an applied research question.
- Theories are built on *concepts*. Constructs are abstract concepts that are not observable but are inferred by measuring relevant or correlated behaviors, such as function or pain.
- A proposition is a generalized statement that asserts the theoretical linkages between concepts.

- A model is a simplified approximation of a process or structure, including conceptual, physical, and statistical models.
- Deductive reasoning is the acceptance of general proposition and subsequent inferences that can be drawn. Deductive theories are intuitive, providing insight that can then be tested.
- Inductive reasoning is logic that develops generalizations from specific observations. Inductive theories evolve through a process that begins with empirical observations that are studied for patterns to form generalizations.
- Theories are important foundations for asking good clinical questions, providing the rationale for interpretation of outcomes.
- Middle-range theories form a bridge between theory and empirical observations, providing opportunities for hypothesis testing under clinical conditions. Grand theories are more comprehensive, trying to explain phenomena at the societal level. Meta-theories are used to reconcile several theoretical perspectives in the explanation of sociological, psychological, or physiological phenomena. Laws are derived when theories reach the level of absolute consistency, mostly observed in physical sciences.

CHAPTER

5

Understanding Evidence-Based Practice

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the factors that have influenced the need for EBP in your profession.
2. Describe sources of knowledge and how they relate to the use of evidence in practice.
3. Define EBP and how the model contributes to clinical decision-making.
4. Describe the five steps in the EBP process.
5. Develop clinical questions using the PICO format for studies of interventions, diagnosis and prognosis.
6. Give examples of background and foreground questions related to a patient case.
7. Describe the general questions used to critically appraise a study.
8. Describe the levels of evidence used to distinguish the strength of studies for quantitative and qualitative studies.
9. Define the importance of implementation studies and knowledge translation to the EBP process.

■ Key Terms

Evidence-based practice (EBP)
Background question
Foreground question
PICO
Systematic reviews
Meta-analysis
Clinical practice guidelines
Scoping reviews
Critical appraisal
Levels of evidence
Knowledge translation (KT)

■ Chapter Summary

- Evidence-based practice (EBP) is the use of best research evidence in conjunction with clinical expertise, patient values, and clinical circumstances, to inform clinical decisions.
- EBP is important because of the gaps in clinical knowledge, and the overuse, underuse, or misuse of available procedures.

- EBP contributes to decision-making, rather than the typical sources of knowledge that include reliance on traditional methods or authorities or experience.
- EBP does not mean that everyone has to use a "cookbook" approach but is part of a larger process that includes experience and judgment within the framework of patient needs and preferences.
- The five steps of EBP include 1) asking a clinical question, 2) acquiring relevant literature, 3) appraising the literature, 4) applying findings to a clinical decision, and 5) assessing the success of the process. These steps are sometimes referred to as the five "A's."
- A background question is related to etiology or general knowledge about a patient's condition, referring to the cause of a disease or condition, its natural history, signs and symptoms, or the anatomic or physiological mechanisms that relate to pathophysiology.
- A foreground question focuses on specific knowledge to inform decisions about patient management. It is phrased using the PICO format: **P**opulation, **I**ntervention, **C**omparison, and **O**utcomes. These questions can relate to diagnosis, measurement, prognosis, intervention, or patient experiences.
- Systematic reviews and meta-analyses are syntheses of literature that provide summaries of evidence to aid in decision-making. They assess the availability of evidence and appraise the quality of the research studies, providing summary conclusions.
- Critical appraisal of literature should address three main questions: Is the study valid? Are the results meaningful? Are the results relevant to my patient?
- Research studies have been classified according to levels of evidence that represent the expected rigor of the study design, indicating the level of confidence that can be placed in the findings. For studies of diagnostic accuracy or prognosis, observational studies will provide the strongest evidence. For studies of interventions, screening, or assessment of harm, RCTs are considered most effective. Systematic reviews should provide the highest level of evidence because they include multiple studies and quality assessment.
- Qualitative research does not fit neatly into the classification system because of its intent to understand human experience. Evidence can be judged according to generalizability of the study based on theoretical premise and ability to extend findings into broader contexts.
- Many barriers exist to the implementation of EBP within clinical settings, including a lack of skill in critical appraisal of literature, the ability to access literature, and the lack of time and resources that are typical in busy clinical environments. Many strategies can be used to establish an evidence-based culture.
- Knowledge translation (KT) is related to the long-standing problem of underutilization of evidence. Although research may demonstrate treatment effectiveness or the presence of risk factors, KT involves the adaptation of quality research into relevant priorities, including the creation and application of knowledge.

CHAPTER

6

Searching the Literature

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Apply strategies for using search engines and databases for locating research literature.
2. Develop search strategies using Boolean logic and Medical Subject Headings.
3. Describe several methods for refining or broadening a search.
4. Distinguish between primary and secondary sources of information.

■ Key Terms

Peer review
Grey literature
Publication bias
Primary source
Secondary source
Database
MEDLINE
PubMed
Digital object identifier (DOI)
Cumulative Index to Nursing and Allied Health Literature (CINAHL)
Cochrane Database of Systematic Reviews
Search engines
Keywords
Truncation
Wildcards
Boolean logic
Nesting
Medical subject headings (MeSH)
Exploding
Focusing
Limits
Filters
Clinical queries
Sensitivity
Specificity
Abstracts
Interlibrary loan
Open access
My NCBI
Citation management applications
Review of literature

Chapter Summary

- Scientific literature includes all scholarly products, including original research articles, editorials, position papers, reviews and meta-analyses, books, dissertations, conference proceedings and website materials.
- Many journals are peer reviewed**,** which means that manuscripts are scrutinized by experts before they are accepted for publication.
- Grey literature is material that is not produced by commercial publishers, including government documents, reports of all types, fact sheets, practice guidelines, conference proceedings and other non-journal products. These can provide important information for evidence-based practice (EBP).
- Primary sources are reports provided directly by the investigator, such as journal articles. Secondary sources include reviews of studies presented by someone other than the original author. Caution must be exercised when using secondary sources, as they may not accurately reflect the material in the original source.
- Systematic reviews and meta-analyses include critical analysis of published works. Although technically a secondary source, these are a form of research in generating new knowledge based on rigorous and documented analysis of previous research.
- Many databases and search engines can be used to access citations and full text of literature. *MEDLINE* is the most comprehensive database of health-related research and can be freely accessed through *PubMed*. Other important databases include *CINAHL* and the *Cochrane Database of Systematic Reviews***.**
- The search process begins by identifying *keywords***.** These terms can be identified within a PICO question. Boolean logic is used to create combinations of keywords that can help to fine tune a search using the terms AND, OR, and NOT.
- Medical Subject Headings (MeSH) are subject headings developed by the National Library of Medicine to sort through keywords.
- Several search strategies can be used to refine a search by truncating keywords, filtering for dates or other characteristics, and accessing *clinical queries* through PubMed to target types of research. Other strategies include finding references in other articles and looking for related articles within a database.
- Reviewing *abstracts* helps to screen for articles that are relevant to your clinical question.
- Full text of articles may be available for free through PubMed and other databases. When free articles are not available, access can be obtained through *interlibrary loan*, author requests, and public libraries. Many journals are open access**.**
- PubMed offers many options that allow you to save searches, to build collections of related citations, and to get e-mail alerts when relevant articles are published.
- *Citation management applications* allow you to save references and organize your citations when writing your papers. The most popular are EndNotes and RefWorks, both of

which require subscription or access through your institution. Several free reference tools are available.

- When completing a *review of literature*, it is helpful to start with a systematic review that can give you an idea of the scope of prior research.

CHAPTER 7

Ethical Issues in Clinical Research

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the major principles in the *Common Rule* and recent changes.
2. Explain the role of HIPAA in clinical research.
3. Define the principles of beneficence, justice, and autonomy as they relate to research ethics.
4. Discuss ethical issues related to use of control groups in human studies research.
5. Discuss historical foundations of research ethics.
6. Describe the role of the institutional review board in clinical research.
7. Describe the elements of informed consent.
8. Describe the process of developing and submitting a proposal to an IRB.
9. Define the three main types of misconduct in research.

■ Key Terms

Nuremberg Code
Declaration of Helsinki
National Research Act
Institutional Review Board (IRB)
Belmont Report
The Common Rule
Respect for persons
Autonomy
Beneficence
Justice
Risk-benefit ratio
Expedited review
Exempt review
Informed consent
Risk
Equipoise
Benefits
Fabrication
Falsification
Plagiarism
Conflict of interest (COI)
Peer review
Retractions
Replication studies

Chapter Summary

- Several documents have become part of ethical standards for human research. The *Belmont Report* codifies three basic principles that guide clinical research.
  - *Respect for persons* involves attention to human dignity, and an individual's right to autonomy or self-determination. This means that individuals have the right to make decisions about their lives, including whether to participate in research studies.
  - Beneficence refers to the obligation to attend to the well-being of individuals, including physical and psychological risks.
  - Justice refers to fairness in the research process, or the equitable distribution of the benefits and burdens. This relates to selection of participants, with the intent that subjects represent diverse elements of the population and that the burden of research does not fall only on disadvantaged groups. It also requires attention to the risk-benefit ratio, to ensure that potential benefits outweigh potential risks.
- Federal regulations require the review of research proposals by an Institutional Review Board (IRB) to ensure the rights and protections of all participants. Proposals may be reviewed fully or may receive *expedited reviews* depending on the nature of the study and potential risks to subjects.
- Informed consent must be obtained from all participants in research. Informed consent forms document the full scope of a study, potential risks and benefits, and protections such as anonymity and confidentiality. These forms must be signed to document fully voluntary participation. Most institutions have specific requirements for drafting consent forms, which will be scrutinized by the IRB. Informed consent forms must include certain elements, which are codified in the Common Rule (see the Chapter 7 Supplement).
- Research integrity on the part of the researcher is an important concern in human studies, with many unfortunate examples of violations of ethics. The three forms of misconduct include fabrication of results, plagiarism, and falsification of data. When studies are found to violate these ethical standards, they may be *retracted* so that readers will know that the findings should not be considered valid.
- *Conflicts of interest (COI)* occur when a researcher's professional judgement may be influenced by secondary personal interests, such as financial gain. Potential conflicts must be disclosed as part of the planning and reporting of research.

CHAPTER 8

Principles of Measurement

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Distinguish among continuous, discrete, and dichotomous variables.
2. Discuss the challenge of measuring constructs.
3. Define and provide examples of the four scales of measurement.
4. Discuss the relevance of identifying measurement scales for statistical analysis.

■ Key Terms

Dichotomous variable
Polytomous variable
Continuous variable
Discrete variable
Precision
Constructs
Levels of measurement
Nominal scale
Ordinal scale
Interval scale
Ratio scale
Parametric tests
Nonparametric tests

■ Chapter Summary

- Measurements are taken to describe quantity, set criteria for performance, make comparisons, evaluate patient conditions, discriminate between individuals with different characteristics, and to predict outcomes or relationships.
- Measured variables can be dichotomous, having only two possible values, or polytomous, having more than two values.
- A continuous variable can take on any value along a continuum, whereas a discrete variable can be described only in whole integer units.
- A construct is an abstract variable that is not observable and is defined by the measurement used to assess it. It is also considered a latent trait because it reflects a property within a person and is not externally observable. Examples are intelligence, health, pain, mobility, and depression.

- Four levels of measurement include:
  - Nominal scale: classifies objects or people into categories with no quantitative order. Example: blood type.
  - Ordinal scale: a rank-ordered measure where intervals between values are unknown and likely unequal. Example: numerical rating scale for pain.
  - Interval scale: possesses rank order but also has known and equal intervals between consecutive values but no true zero. Example: temperature.
  - Ratio scale: the highest level of measurement is an interval scale with an absolute zero point. Example: height and weight.
- Nominal values can only be counted. Ordinal measures can be ranked and expressed as counts or percentages. Technically these values should not be subjected to arithmetic operations, but they often are. Mathematical calculations can be used with interval and ratio data.
- Parametric statistics, which are designed to estimate population values, require interval or ratio data. Nonparametric statistics are used with ordinal or nominal measures. There are conditions when ordinal values may also be used with parametric statistics.

CHAPTER 9

Concepts of Measurement Reliability

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the importance of reliability in clinical measurement.
2. Define reliability in terms of measurement error.
3. Distinguish between random and systematic error.
4. Describe typical sources of measurement error.
5. Describe the effect of regression toward the mean in repeated measurement.
6. Discuss how concepts of agreement and correlation relate to reliability.
7. Define and provide examples of test-retest, rater, and alternate forms reliability.
8. Discuss how generalizability theory influences the interpretation of reliability.
9. Discuss how reliability is related to the concept of minimal detectable change.

■ Key Terms

Reliability
Classical measurement theory
True score
Systematic error
Random error
Reliability coefficient
Relative reliability
Absolute reliability
Classical measurement theory
True score
Systematic error
Random error
Reliability coefficient
Relative reliability
Absolute reliability
Intraclass correlation coefficient (ICC)
Standard error of measurement (SEM)
Generalizability theory
Test-retest reliability
Carryover
Testing effects
Intra-rater reliability
Inter-rater reliability
Alternate forms reliability
Internal consistency
Split-half reliability
Change score

Regression toward the mean (RTM)
Minimal detectable change (MDC)

Methodological research

Chapter Summary

- In classical measurement theory an observed score consists of two components: a true score which is a fixed value, and unknown error.
- Systematic error is predictable, occurring in a consistent overestimate or underestimate of a measure. Random error refers to errors that have no systematic bias and can occur in any direction or amount.
- Errors of measurement can occur because: 1) the individual taking the measurement (the rater) does not perform the test properly; 2) the measuring instrument itself does not perform in the same way each time; or 3) the variable being measured is not consistent over time.
- Reliability is an indicator of the degree to which a measurement contains error. Relative reliability is expressed as a coefficient, which reflects how much of the total variance in a set of scores is true variance, and not error. Reliability coefficients can range from 0.0 to ±1.00. Consideration of what coefficient value is considered acceptable reliability must be made within the context of the measurement and the degree of precision needed for making judgments about the measurement.
- Absolute reliability indicates a degree of error variance in the actual units of a measurement, reflecting the measured amount of error that can be expected.
- Reliability exists to some extent in every measuring instrument and is not an all-or-none trait.
- Generalizability theory suggests that not all error is the same, and can be partitioned to understand specific sources, such as rater or timing.
- Test-retest reliability is an assessment of how well an instrument will perform from one trial to another when the actual measurement has not really changed.
- Intra-rater reliability is a measure of the stability of data recorded by one tester across two or more trials. Inter-rater reliability concerns variation between two or more raters who are measuring the same property.
- Alternate forms reliability is used to assess whether alternative versions of a test can provide reliable scores. This may be in the form of a test or questionnaire, or it may be comparison of different models of an instrument.
- Internal consistency reflects the extent to which items on a multi-item inventory are measuring the same characteristics.
- Change scores reflect the difference in performance from one session to another, often a pretest and posttest. If measures do not have strong reliability, change scores may primarily be a reflection of error.
- The phenomenon of regression toward the mean indicates that extreme scores on an initial test are expected to move closer to the group average on a second test, even when there is no true change in performance. This phenomenon is reduced if a measurement has strong reliability.
- Minimal detectable change (MDC) is the amount of change in a variable that must be achieved beyond the minimal error in a measurement. It is a threshold above which

we can be confident that a change reflects true change and not just error.

- Methodological research involves the development and testing of measuring instruments to establish reliability.
- Methods for improving reliability include standardizing measurement protocols, training raters, calibrating instruments, and taking multiple measurements.

CHAPTER 10

Concepts of Measurement Validity

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the importance of validity in clinical measurement.
2. Define and provide examples of face, content, criterion-related, and construct validity.
3. Discuss issues affecting validity of measuring change.
4. Define minimal clinically important change.
5. Distinguish between criterion and norm referencing.

■ Key Terms

Validity
Content validity
Criterion-related validity
Concurrent validity
Predictive validity
Construct validity
Convergent validity
Discriminant validity
Face validity
Gold standard
Sensitivity
Specificity
Known groups method
Multitrait-multimethod matrix (MTMM)
Factor analysis
Exploratory factor analysis (EFA)
Confirmatory factor analysis (CFA)
Criterion-referenced test
Norm-reference test
Responsiveness
Minimal detectable change (MDC)
Minimal clinically important change (MCID)
Floor effect
Ceiling effect

Chapter Summary

- Validity concerns the meaning that we give to a measurement, reflecting if a measure is capable of discriminating among individuals with and without certain traits, if it can evaluate the magnitude or degree of change in a measurement, and if we can use it to make predictions about future status based on a measurement.
- Validity is affected by systematic error.
- Validity of physical measures is straightforward, such as measurement of length using a ruler. Measuring constructs presents a challenge to assess validity.
- Validity is not an all-or-none property and will be present to some degree in most measures. Therefore, rather than asking if an instrument is valid, it is more appropriate to ask if it is valid for a given purpose.
- Measuring validity is not as straightforward as reliability. It must be evaluated by constructing an "evidentiary chain" that links the method of measurement to the intended application. Three main types of evidence can support validity.
- Content validity establishes that the content of a test adequately samples the universe of content that defines the construct being measured. This is typically established through consensus of experts.
- Criterion-related validity establishes the correspondence between a target test and a reference standard to determine that the target test is measuring the variable of interest. Concurrent validity establishes this correspondence by taking both measurements at relatively the same time. Predictive validity reflects the extent to which the target test can predict a future reference standard.
- The standard used to establish criterion validity is ideally a gold standard that is already established as a valid measure of the variable of interest. When a gold standard is not available, as is true for many clinical variables, a reference standard must be chosen that is assumed to measure the variable of interest. This is the process used in the assessment of diagnostic accuracy and screening tests.
- Sensitivity indicates the extent to which the target test accurately identifies those with the condition (true positive), and specificity measures the ability of the target test to identify those without the condition (true negatives).
- Construct validity establishes the ability of an instrument to measure the dimensions and theoretical foundations of an abstract construct.
- Methods of establishing construct validity include the known groups method whereby individuals known to be different on a variable are tested to determine that the instrument can identify their group membership. Factor analysis is a statistical procedure that examines multiple dimensions of a scale to determine if the theoretical components of the measurement are adequately represented.
- Convergent validity measures the extent to which a test correlates with other tests of similar constructs, and discriminant validity is the extent to which a test is not correlated with tests of different constructs.
- Face validity is not considered a true measure of validity but does reflect the degree to which an instrument appears to measure what it is intended to measure.
- Criterion referencing refers to comparison of an individual's performance on a test

to a fixed standard that represents an acceptable level of behavior, such as a passing grade on an exam or measurement of a normal heart rate. Norm-referencing is used to compare and rank individuals within a defined population, indicating how their performance ranks within the group, such as a curved grade on an exam.

- The minimal clinically important difference (MCID) is the smallest difference in a measured variable that signifies a meaningful difference in performance. It is based on a criterion that established sufficient change to be considered "important." This may be a subjective determination by a patient ("I feel better"), or it may reflect a clinician or family member's view of improvement.
- A ceiling effect occurs when a scale is not precise enough at the high end to distinguish small changes in those with strong performance. A floor effect occurs when the scale cannot sufficiently distinguish small changes at the bottom of the scale.
- Methodological research for validity incorporates testing both reliability and validity in several ways over time. Strategies for improving validity include fully understanding the construct of interest and its clinical context, using several different approaches for gathering evidence of validity, and cross-validating outcomes on different groups to show consistency in outcomes.

CHAPTER 11

Designing Surveys and Questionnaires

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the role of surveys in clinical research.
2. Describe the basic structure of survey instruments.
3. Describe the process of designing a survey.
4. Discuss the characteristics of good survey questions.
5. Develop a cover letter that will engage potential participants in a survey.
6. Describe the processes of Q sort and Delphi surveys.

■ Key Terms

Survey
Questionnaire
Interviews
Structured interview
Semi-structured interview
Self-report
Recall bias
Open-ended question
Closed-ended question
Likert scale
Visual analog scale (VAS)
Branching
Population
Sample
Probability sample
Nonprobability sample
Sampling error
Response rate
Cross-tabulations

■ Chapter Summary

- Surveys can be used for descriptive purposes or to generate data to test specific hypotheses.
- Surveys are generally administered as questionnaires or through interviews. Interviews can be structured, semi-structured, or unstructured, depending on the format of questions.
- Because surveys are generally based on self-report, recall bias can be an issue if respondents do not accurately remember behaviors.

- Stage 1 in planning a survey involves delineation of a research question, the target population, and hypotheses. Stage 2 includes designing the survey instrument, creating and testing items, and finalizing the format. Stage 3 includes IRB approval, selection of a sample, and distribution of the survey.
- Like any other form of data collection, surveys are part of a larger research question that specifies the variables of interest.
- Developing content for a survey involves reviewing existing instruments to determine if they are applicable for your study. If items must be developed, they must go through several rounds of review, usually by experts in the field. Through pilot testing and revisions, the content if refined until it attains an acceptable level of validity.
- Survey questions can be open-ended, where respondents are asked to answer in their own words, or they may be closed-ended, where respondents are asked to choose from a fixed set of choices for their answer.
- There are many types of closed-ended questions, including dichotomous choices, check all that apply, multiple choice, measures of intensity, checklists, or scales.
- Writing good questions requires using clear language, giving applicable answer choices, and considering if respondents can answer easily and honestly.
- Questions should be worded clearly, avoiding bias in wording, and using language that can be easily understood by all respondents.
- Samples for surveys generally need to be large. Probability samples are preferable but may not be feasible.
- Coverage error occurs when all members of a population are not eligible, potentially biasing results.
- Response rates for surveys are the major drawback of this method. Needed sample size can be determined based on estimates of an acceptable margin of error.
- The cover letter is an important part of the survey to give respondents the motivation to respond by letting them know the importance of the project and why they were chosen. Confidentiality should be protected. Follow up is often necessary to encourage respondents to return the survey.
- Questions and responses should be coded so that data can be easily recorded once surveys are returned.
- Data are often subjected to descriptive statistics or cross-tabulations to determine if answers are related.
- Returning the survey is considered informed consent.

CHAPTER 12

Understanding Health Measurement Scales

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe how measurement scales are used to evaluate constructs.
2. Distinguish between summative and cumulative scales.
3. Describe the characteristics of a Likert scale.
4. Discuss the application and limitations of visual analog scales in clinical practice.
5. Describe the role of Rasch analysis for understanding cumulative scales.

■ Key Terms

Scale
Subscales
Summative scale
Cutoff scores
Likert scale
Visual analog scale (VAS)
Anchors
Numerical rating scale
Faces rating scale
Cumulative scale
Classical measurement theory
Item response theory
Rasch analysis
Person-item map
Logit
Thresholds
Item distribution
Item separation
Person separation
Infit
Outfit
Person fit
Differential item functioning (DIF)
Computer adaptive testing (CAT)

■ Chapter Summary

- Many measurement scales, usually designed as questionnaires, are composed of several items that are used to reflect an underlying latent trait and the dimensions that define it.
- A scale is an ordered system based on a series of questions or items, resulting in a score that represents the degree to which a respondent possesses a particular attitude,

value, or characteristic. In classical measurement theory, the trait is assessed by getting a total numerical score.

- For a scale to be meaningful, it should be unidimensional, representing a singular overall construct. Subscales can be created if different dimensions exist with the overall trait.
- Scales may be generic or they may be geared towards specific conditions, age ranges, or care settings.
- A summative scale is one that presents a total score by adding values across a set of items. The score is intended to indicate the "amount" of the latent trait that is present in an individual. The summative score is based on the assumption that all items contribute equally to the total.
- The interpretation of summative scales can be problematic because the same score can be achieved by two people, but their actual trait intensity may be different if they have a different combination of responses.
- Many scales use items with scores on an ordinal scale. These can present interpretation difficulties because the intervals between graded scores are not equal or known. When scales are used as screening tools, a cutoff score can be determined to demarcate a positive or negative test.
- A Likert scale is a summative scale in which individuals choose among answer options that range from one extreme to another to reflect attitudes, beliefs, perceptions or values. They typically include five categories, such as: Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), Strongly Agree (SA). Four categories can be used, eliminating the neutral choice, to force respondents to declare their opinion.
- The Multidimensional Health Locus of Control (MHLC) scale is an example of a Likert scale that includes three subscales indicating if the individual believes controlling health is primarily internal, a matter of chance, or under the control of powerful others.
- A visual analog scale (VAS) is composed of a line, usually fixed at 100 mm in length, with word anchors on either end that represent extremes of a characteristic, such as "no pain" and "worst pain imaginable" or "fatigue" and "worst possible fatigue." The individual is asked to place a mark along the continuum that represents the level of the trait they experience. This scale is considered a quick self-report measure. It can only measure a trait in one dimension, although multiple VASs can be used to assess multiple dimensions.
- Numerical rating scales ask for a respondent to indicate the extent of their pain (or other trait), usually from 0 to 10, with anchors defined at each extreme. A faces rating scale has been used with faces showing increments of distress, often used with children or elderly individuals.
- VAS scores, although recorded in millimeters, are ordinal, as it is not possible to know if one individual's score of 6 is equal to another's score at the same level.
- A cumulative scale presents a set of statements that reflect increasing severity of the characteristic being measured. It is based on the assumption that there is only one unique combination of responses that can achieve a particular score, as each item is required to be present for the next item to be selected.
- In Rasch analysis, items are examined to indicate which ones are more or less difficult, and persons' responses are examined to reflect their level of ability. Based on item response theory (IRT), this model will indicate a good fitting scale if those individuals with greater disability can "pass" only the easier items, and those with greater ability can "pass" the more difficult items.

- A good scale will have items that reflect the full range of difficulty as well as the full range of ability and will have items that are distributed across that continuum. Fit statistics are used to indicate how well the scale fits this model. Because the relationship between difficulty and skill level should be constant, a given score should represent the same degree of a trait for all persons with that score.
- Differential item functioning (DIF) refers to potential item bias in the fit of the data when subgroups within a sample respond differently to individual items, despite their having similar levels of the latent trait.
- Computer adaptive testing (CAT) involves adjusting levels of difficulty of items on a test based on how individuals respond to questions. Using a complex algorithm examining correct and incorrect responses, the number and difficulty of succeeding items can be altered so easier items can be omitted for those with higher performance.

CHAPTER

13

Choosing a Sample

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Distinguish between populations and samples.
2. Define the concepts of sampling bias and sampling error.
3. Distinguish between target and accessible populations.
4. Describe the purpose of inclusion and exclusion criteria in sampling for research studies.
5. Contrast different types of probability and nonprobability sampling procedures.
6. Discuss issues in recruiting an adequate sample size.

■ Key Terms

Population
Sample
Target population
Accessible population
Inclusion criteria
Exclusion criteria
Power
Probability sample
Random sampling (selection)
Nonprobability sample
Statistic
Parameter
Sampling error
Sampling bias
Random sample
Simple random sample
Systematic sample
Stratified random sample
Cluster sampling
Multi-stage sampling
Disproportional sample
Convenience sample
Consecutive sample
Quota sampling
Purposive sampling
Snowball sampling

Chapter Summary

- A population is the aggregate of persons or objects that meet a specified set of criteria and to whom results of a study will be generalized.
- A sample is a subgroup of the population chosen for study to serve as the reference group for drawing conclusions about the population.
- The target population is the overall group to which results will be generalized. The accessible population is the group to which the researcher has access and from which the actual sample will be drawn.
- When selecting a sample, the researcher must specify inclusion criteria that identify the primary traits that would make someone from the population eligible to be a sample subject, such as demographics, education, or health conditions. Exclusion criteria are those factors that would preclude someone from being a subject, factors that could confound results such as comorbidities or cognitive ability.
- A recruitment strategy has to be developed that will indicate how individuals will be identified from the accessible population and how they will be invited to participate.
- The final sample will be composed of those who agree to participate. The sample may be further reduced by the end of the study because of attrition. Reasons for attrition should be documented to determine if they are related to the study, such as participants who drop out because of inconvenience, or if reasons are random.
- A flow diagram should be included that details the number of participants who are invited, who agree to participate, and who are part of the final analysis, with reasons for attrition.
- The power of a study relates to the ability to find statistically significant effects when they exist. Factors that affect power include sample size, variability in the data, and the magnitude of the effect being studied. Therefore, sample size should be determined during planning of a study to be sure that it is sufficient to identify outcomes.
- Samples can be identified through a process of probability sampling, in which participants are randomly chosen from the accessible population, meaning that each member has an equal chance of being selected, eliminating potential bias. In a nonprobability sample, participants are chosen by nonrandom methods. This technique is used more often in clinical research because of the availability of individuals to serve as subjects.
- Sampling error refers to the difference between a sample's averaged data and the averages that would be obtained from the entire population. Although we can't know this difference, there are ways to estimate it statistically.
- Sampling bias occurs when the individuals selected for a sample overrepresent or underrepresent certain population attributes that are related to the phenomenon under study. Such bias should be minimized when a random sample is chosen.
- Probability sampling methods include several randomized approaches.
  - Simple random sampling involves choosing subject at random from the accessible population, usually based on directories or census lists.
  - Systematic sampling draws participants from lists by choosing individuals based on a sampling interval, such as every 10^{th} person.

- Stratified random sampling is used when the population is divided into subsets based on a characteristic that is relevant to the outcome, such as age or gender. Subjects are chosen randomly from each stratum, usually in proportion to the number of individuals in the population who fall into that subgroup.
- Cluster sampling is used when accessing a large or dispersed population. Individuals or groups are clustered according to some characteristic and randomly chosen within each cluster. For example, a sample may start by randomly choosing 10 states, then 10 cities within each state, and then residents within certain areas in those cities, a technique called multi-stage sampling.
- Disproportional sampling is used when subgroups in the population are of unequal size, creating a situation where some groups may provide insufficient samples for comparison. Smaller groups may be purposefully *oversampled* to give them adequate representation. Analysis procedures are adjusted to account for this disproportional representation.

- Nonprobability samples are created when samples are chosen on some basis other than random selection.
  - In convenience sampling, subjects are chosen based on availability, including the use of volunteers.
  - Consecutive sampling involves recruiting participants as they become available.
  - Quota sampling incorporates elements of stratification so that participants are chosen nonrandomly from subgroups.
  - In purposive sampling, subjects are chosen based on particular criteria and asked to participate, in order to create an informative sample for a specific purpose.
  - Snowball sampling is used when individuals are difficult to identify, such as in the study of sensitive topics or rare characteristics. A few subjects are identified who meet the criteria and they are asked to identify others who they know who also fit the requirements. This "snowballing" is continued until an adequate sample is obtained.

CHAPTER

14

Principles of Clinical Trials

Chapter Overview

■ Objectives

1. Describe the role of clinical trials in healthcare research.
2. Describe the characteristics of a randomized controlled trial.
3. Define different random assignment strategies.
4. Describe different types of control groups.
5. Discuss the importance of blinding in research protocols.
6. Discuss the differences between randomized controlled trials and pragmatic clinical trials.
7. Define phases of clinical trials.
8. Discuss the difference between superiority and non-inferiority trials.

■ Key Terms

Clinical trial
Therapeutic trial
Diagnostic trial
Preventive trial
Clinical trials registry
Randomized controlled trial (RCT)
Treatment arms
Parallel group
PICO
Active variable
Attribute variable
Random assignment (allocation)
Simple random assignment
Block random assignment
Stratified random assignment
Cluster random assignment
Random consent design
Assignment by patient preference
Run-in period
Allocation concealment
Control group
Placebo
Sham
Attention control group
Wait list control group
Active controls
Equipoise
Blinding
Double-blind study
Single-blind study
Open-label trial
Pragmatic clinical trial (PCT)
Pragmatic-explanatory continuum
Preclinical research (basic research)
Phase I trial
Phase II trial
Phase III trial

Phase IV trial
Superiority trial
Non-inferiority trial
Non-inferiority margin
Minimal clinically important difference (MCID)
Equivalence trial

Chapter Summary

- Clinical trials can focus on therapeutic effects of interventions, establishing accuracy of diagnostic tests, or prevention strategies.
- The randomized clinical trial (RCT) is considered the gold standard for experimental research based on the rigor of its structure in controlling confounding effects.
- The basic RCT design incorporates two randomly assigned groups. Each group is also called a treatment arm.
- Experimental designs require the ability to define the levels of an independent variable, the use of random assignment to groups, and the use of a control group for comparison.
- An independent variable can be an active variable when subjects can be assigned to groups or conditions. An attribute variable is a factor that cannot be assigned because it is an inherent characteristic of participants.
- Random assignment is a foundational element of an RCT, meaning that each subject has an equal chance of being assigned to any group, creating groups that should be balanced.
- There are several forms of random assignment that can be used.
  - In simple random assignment, every member of the sample has an equal chance of being assigned to either group.
  - In block random assignment, subjects are randomly divided into even-numbered subgroups (blocks) and are randomly assigned to treatment arms within each block.
  - Stratified random assignment involves subjects being divided into strata based on a relevant group characteristic, and then randomly assigned to treatment groups within each stratum.
  - Cluster random assignment involves assigning participants to groups based on sites in which they are available, randomly assigning treatments to sites, and all individuals at that site getting the same treatment.
  - The random consent design involves assigning subjects randomly to groups before seeking consent to participate. Only those assigned to the experimental group are approached for consent, and others receive standard care.
  - Assignment by patient preference allows subjects to express a preference for one treatment condition or the other. Only those who express no preference are randomly assigned to groups.
  - A run-in period is a way of assuring that subjects will adhere to the protocol. All subjects are given a placebo prior to assignment, and those who comply are then randomly assigned to be part of the study.
- Allocation concealment involves assignment of subjects to groups without the knowledge of those involved in the experimental process. A common method is use of sealed envelopes that contain a

participant's assignment. External agencies can generate random sequences to assure that there is no bias.

- Using a control group is an essential design strategy in an RCT providing a comparison for a treatment group. The control can be a standard treatment, a placebo or sham, or no intervention at all.
- Blinding, also called masking, assures that those involved in the study are unaware of the participant's group assignment. In a double-blind study, neither subjects nor investigators are aware of treatment groups. In a single-blind study, either subjects or investigators may be blinded, but not both.
- In open label trials, both researchers and participants know which treatment is being administered, which is often necessitated by the nature of the treatment.
- RCTs represent the gold standard for experimental research, providing the strongest level of control. However, the need to define samples narrowly and define exact protocols can make generalization to the real world difficult. **A** pragmatic clinical trial (PCT) is designed like an RCT but incorporates a diverse patient population recruited directly from practice settings. Treatment proceeds as it would in typical clinical situations and data collection focuses on important outcomes such as patient satisfaction and quality of life.
- Randomized trials are defined in sequential phases related to the types of information collected, usually related to drug trials. In Phase I trials, small groups of healthy volunteers are tested to demonstrate safety of a new therapy. In Phase II trials, larger groups of patients are evaluated to determine the efficacy of the treatment at different dosages. In Phase III trials, very large samples are tested over long periods to compare the new treatment with standard care, monitoring effects of dosages and side effects. Phase IV trials are done after new treatments are approved. Subgroups may be tested to explore the effects of the treatment further.
- When studies are done specifically to show that one treatment is superior to another, the trial is considered a superiority trial**.** When a new treatment is being compared to an accepted treatment to show that it is equally effective, the trial is considered a non-inferiority trial, with the intent of showing that the new treatment is a reasonable alternative, which may be cheaper or have fewer risks. The new treatment is then considered "no worse" than the standard approach. The difference between the treatments must be within a non-inferiority margin that is the biggest difference between the responses of the two groups that would be acceptable to consider the new treatment an acceptable substitute.

CHAPTER 15

Design Validity

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe threats to internal, external, construct and statistical conclusion validity related to quantitative research.
2. Describe several strategies to control for subject variability in experimental design.
3. Describe the purpose of blinding.
4. Describe strategies for controlling intersubject differences in research design.
5. Discuss various reasons for missing data and different types of "missingness."
6. Describe how a flow diagram can explain how a sample is finalized.
7. Explain the purpose of intention to treat analysis.
8. Describe various methods for handling missing data in data analysis including data imputation.

■ Key Terms

Statistical conclusion validity
Internal validity
Internal threats
History
Maturation
Attrition
Testing effect
Reactive effect
Instrumentation effect
Selection
Quasi-experiment
Social threats
Diffusion or imitation of treatments
Compensatory equalization
Operational definitions
Multiple treatment interaction
Experimental bias
Experimenter effect
External validity
Adherence
Hawthorne effect
Ecological validity
Homogeneous sample
Blocking variable
Matching
Propensity score

Independent factor
Repeated factor (measure)
Analysis of covariance (ANCOVA)
Covariate
Per-protocol analysis
Intention to treat (ITT)
Missing completely at random (MCAR)
Missing at random (MAR)
Missing not at random (MNAR)
Completer analysis
Imputation
Last observation carried forward (LOCF)
Multiple imputation

■ Chapter Summary

- Whether using explanatory or observational designs, four types of validity are important to understanding the strength of evidence: statistical conclusion validity, internal validity, construct validity, and external validity.
- Statistical conclusion validity concerns the appropriate use of statistical procedures for analyzing data, leading to unbiased conclusions about the relationship between independent and dependent variables. Threats include a lack of statistical power, violated assumptions of statistical tests, reliability of the data, and failure to use intention to treat analysis.
- Internal validity has a focus on cause-and-effect relationships, requiring three components in a design.
  - Temporal precedence means that the cause precedes the effect, that a change in outcome is only observed following the application of treatment.
  - Covariation of cause and effect is based on documentation that the outcome only occurs in the presence of the intervention and therefore is not related to other causes.
  - Finally, we must demonstrate that *no plausible alternative explanation***s** are possible and that confounding variables are not responsible for the outcome rather than the intervention.
- Specific threats to internal validity include:
  - History refers to the confounding effect of specific events that occur after the introduction of treatment but before the final test.
  - Maturation concerns the effect of the passage of time on changes in outcomes.
  - Attrition is defined as loss of subjects over time, which is a concern if it is too large or if the attrition is related to the experimental situation.
  - Testing effects occur when the actual testing of the outcome variable is responsible for changes in performance.
  - Instrumentation effects occur when the measuring instruments are unreliable.
  - Regression to the mean is also associated with a lack of reliability, occurring when there are extreme scores on the pretest, which often regress toward the mean score on the posttest.
  - Selection effects occur when subjects are assigned to groups in a biased way, thereby influencing how groups

respond.

- Social threats to validity include diffusion of effects because of communication among subjects, investigators giving preference to one group, or subjects feeling more or less invested because of the receiving the control condition instead of the experimental treatment.
- Construct validity concerns how the independent and dependent variables are defined and is related to the nature of conclusions that can be drawn based on those operational definitions. This can be affected by the time frame of the study, interactions of multiple treatments, or experimental bias that occurs when experimenters or subjects try harder because they are in a study.
- External validity concerns the extent to which results can be generalized beyond the study sample and the experimental setting. It can be influenced by how subjects are selected.
- Strategies to control threats to validity include:
  - Using random assignment to groups
  - Using a homogeneous sample, **r**estricting subjects on certain variables to avoid confounding.
  - Building in a blocking variable as another independent variable to account for it in the analysis.
  - Matching subjects across groups to assure there are comparable subjects.
  - Using repeated measures, where subjects act as their own control, which avoids having to match subjects across groups.
  - Using an analysis of covariance (ANCOVA) as an analysis technique to control for a confounding variable in the analysis of data.
- It is also important to control for the effect of missing values in data, as they can influence analysis, especially if they are not random events. Data may be missing because subjects drop out or if they are noncompliant. A flow diagram should be included in studies to show how subjects proceed through the study, including reasons for subjects who do not complete the study as expected.
- There are many approaches to data analysis to account for missing data. These include per protocol analysis, in which subject data are only included for those who complete the study and complied with the protocol. This can overestimate the effect of a treatment. A preferred method is intention to treat analysis (ITT) which means that data are analyzed according to original random assignment, regardless of the treatment subjects actually received. This approach guards against bias of drop outs.
- When many data points are missing, it could affect data analysis procedures that require all variables to be available. Data can be estimated for missing values using a process of data imputation, which statistically projects values for missing data based on averages or correlations among variables within a sample.

CHAPTER 16

Experimental Designs

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the structure of basic experimental designs for independent groups.
2. Discuss how experimental designs control for threats to internal validity.
3. Discuss the role of main effects and interaction effects in a factorial design.
4. Describe the structure of repeated measures designs.
5. Discuss the advantages and disadvantages of using repeated measures designs.
6. Describe the process of sequential clinical trials.

■ Key Terms

Completely randomized designs
Between-subjects designs
Parallel groups designs
Within-subjects design
Repeated measures design
Single factor design
One-way design
Multi-factor design
Pretest-posttest control group design
Placebo
Posttest-only control group design
Factorial designs
Main effect
Interaction effect
Randomized block design
Blocking variable
Practice effects
Carryover effects
Order effects
Latin square
Crossover design
Washout period
Two-way design
Mixed design
Sequential clinical trials

■ Chapter Summary

- Experimental designs can take on a variety of configurations, all designed to offer control of internal validity.
- A true experiment requires that participants can be randomly assigned to at least two comparison groups. In a one-way design, a study includes one

independent variable. A two-way design includes two independent variables.

- Designs can be chosen based on the number of independent variables, the number of levels within each independent variable, the number of groups being tested, how subjects are assigned to groups, how many observations are made, and the temporal sequence of interventions and measurements.
- The pretest-posttest control group design is the basic structure of a randomized controlled trial (RCT) comparing two groups, also called a parallel group study. Because groups are assigned at random, theoretically they differ solely on the basis of the provided treatment, allowing for conclusions regarding the effect of the treatment.
    - Control groups may receive a placebo, a sham treatment, or a different treatment, typically based on standard care.
    - This design can be extended to include more than two groups to compare several treatment conditions.
- The posttest- only control group design is identical to the pretest-posttest design, except that there is no pretest administered. This approach can be sued when a pretest is either not relevant, impractical, contraindicated, or potentially reactive. It is still considered to have strong control if the groups are randomly allocated.
- When more than one independent variable is incorporated, a factorial design is used. The design is described by the number of independent variables and their levels. Therefore, a 3×2 design has two independent variables, one with 3 levels, the other with 2 levels.
    - This design allows the researcher to study the effect of each independent variable separately, called **main effects**, and the **interaction** between the two variables.
    - In a randomized block design, an attribute variable is built into the design as a second independent variable, allowing its effect to be studied at each level of the treatment.
- In a repeated measures design, subjects are tested at each level of the independent variable, serving as their own controls. This is also called a within-subjects design.
    - A repeated measures study may use a *one-way design* with only one independent variable or a *two-way design* when two independent variables are included.
    - Repeated measures designs must be able to account for the potential carryover or practice effects that can occur when subjects repeat a measurement over several trials, which can be controlled by randomizing the order of conditions.
    - In a cross-over design, half the subjects are assigned treatment one first, the other half getting treatment two. After a washout period, the subjects receive the alternative treatment.
    - A two-way design that includes a repeated measure and an independent measure is called a mixed design**.**
- In a sequential clinical trial, two treatment conditions A and B are compared by analyzing the preferences of pairs of subjects, who are recruited sequentially. The results of

the comparison within each pair is charted on specially constructed charts that provide stopping rules for deciding if treatment A is better than treatment B or if there is no significant difference between them. This design often results in the need for fewer subjects than in a fixed sample design.

CHAPTER 17

Quasi-Experimental Designs

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the reasons for implementing quasi-experimental designs.
2. Describe the basic elements of time-series designs.
3. Describe design variations of time-series designs, including the role of a withdrawal period.
4. Describe nonequivalent group designs
5. Discuss the threats to validity of quasi-experimental designs.

■ Key Terms

Quasi-experimental designs
One-group pretest-posttest design
Interrupted time series design (ITS)
Nonequivalent group designs
Nonequivalent pretest-posttest control group design
Intact groups
Subject preferences
Historical controls
Nonequivalent posttest-only control group design

■ Chapter Summary

- Quasi-experimental designs are similar to experimental designs but lack either random assignment, comparison groups, or both.
- Quasi-experimental designs present threats to internal validity because of the lack of these controls.
- Time series designs focus on assessment of responses over time, where time serves as the independent variable that is measured across several time intervals.
- In the one-group pretest-posttest design, one group of subjects is tested before and after an intervention. The design offers little control of

threats to internal validity. However, the design can be defended if previous research has shown the intervention is better than no treatment, and where the time interval from pretest to posttest is relatively short, limiting the potential effect of other variables on the outcome.

- Time series designs are considered repeated measures, since the same subjects are being tested at time interval.
- The interrupted time series design is an extension of the one-group pretest-posttest design, with several measures before and after an intervention. Data patterns are observed to demonstrate that there is a difference in response following the intervention.
  - Design variations include the use of a continuous intervention that is withdrawn at some point to demonstrate behavior improves with treatment and reverts back to baseline when the treatment is withdrawn.
- The nonequivalent group design can be configured like pretest-posttest experimental designs but without a randomized control.
  - This design can be used with intact groups or when subjects self-select their group assignment.
  - Design validity is threatened by the lack of randomization. Control can be imposed by checking the balance in baseline scores between groups, by stratifying subjects to control for important characteristics, or matching.
  - Historical controls involve the use of subjects from a prior study to serve as controls, compared to subjects receiving the experimental intervention.
  - The nonequivalent posttest-only design includes a control group but is subject to confounding because there is no way to assure equivalence on important characteristics at baseline. This study can be used most effectively to look at relationships among variables for those who do and do not get the intervention, but cause and effect cannot be interpreted.

CHAPTER 18

Single-Subject Designs

Chapter Overview

Objectives

After completing this chapter, the learner will be able to:

1. Describe the basic structure of single-subject designs.
2. Describe methods of measuring target behaviors.
3. Discuss the limitations of A-B designs.
4. Describe the structure of withdrawal designs, multiple baseline designs, alternating treatment designs, and multiple treatment designs.
5. Describe the structure and use of N-of-1 trials.
6. Explain how baselines and withdrawal phases support the internal validity of single-subject designs.
7. Describe the process of visual analysis in single-subject designs.
8. Apply methods of quantitative analysis for single-subject designs.
9. Describe the process of generalization in single-subject research.

Key Terms

Single-subject designs
Target behavior
Baseline phase
Intervention phase
A-B design
Stability
Slope
Trend
Interval recording
Replication of effects
Withdrawal designs
A-B-A design
A-B-A-B design
Multiple baseline design
Alternating treatment design
Multiple treatment design
Changing criterion design
N-of-1 trial
Visual analysis
Level
Trend
Split-middle line
Celeration line
Binomial test
Serial dependency
Two standard deviation band method
Statistical process control (SPC)
Control charts
Effect size
Nonoverlapping scores
Generalization
Direct replication
Systematic replication
Clinical replication

Chapter Summary

- Single-subject designs (SSD) involve serial observations of individual behaviors before, during, and after interventions to document the trend in responses.
- SSDs have the advantage of being able to examine the responses of each individual rather than looking only at aggregated data across groups, contributing to an understanding of patient differences and the benefits of individualized care.
- SSDs are a form of time-series study with a collection of repeated measurements and are therefore considered a within-subjects design.
- The response to treatment is considered the target behavior, which can be measured by frequency, duration, or magnitude of the response.
- Data are collected across at least two phases: A is the baseline during which the target behavior is monitored but no treatment is provided. B is the intervention phase during which the treatment is provided while measurements continue.
  - A design with one baseline and one intervention phase is called an A-B design.
  - Results are plotted on a chart showing the responses across all phases, with the target behavior along the Y-axis, and time across the X-axis usually in hours, days, or weeks.
- Baseline data are an essential element of the SSD, serving as the control condition for comparison with the intervention phase. Baselines may be level, show erratic values, or may be ascending or descending.
  - It is important to demonstrate that the trend in behavior is different in the baseline and intervention phases to document the treatment's effect.
- A-B designs are especially susceptible to threats to internal validity because there is no control condition.
  - Control must be demonstrated through replication of effects, which can involve withdrawal and reinstatement of treatment, replication across more than one subject, or by comparing two or more interventions.
- In an A-B-A withdrawal design, the treatment is withdrawn at some point following the intervention. If responses improve with intervention and then revert back to baseline, it is evidence that the treatment had an effect.
  - Withdrawal designs can also include a second intervention phase, A-B-A-B, to further demonstrate the treatment's effect.
- In a multiple baseline design, effects can be replicated across more than one subject, across multiple settings, or across multiple behaviors.
  - The multiple baseline design can use the A-B configuration because effects can be replicated, controlling potential confounding.
  - Each replication incorporates a different length of the baseline period. By staggering baselines, consistent effects of treatment can be documented.
- In an alternating treatment design, two or more treatment conditions are implemented at each treatment session, usually in random order, to determine if the treatments engender different levels of response.
- In a multiple treatment design, more than one intervention is tested, usually after withdrawal of the initial intervention, using an A-B-A-C configuration.

- Other variations of this design can also include combinations of interventions, such as A-B-A-BC, where B and C interventions are provided alone and in combination to see how their effects differ.

- In a changing criterion design, the treatment goal is incrementally increased to demonstrate the effectiveness of intervention as the patient progresses.
- An N-of-1 trial uses the structure of an RCT to demonstrate the effectiveness of an intervention, or to compare two interventions in a single patient.
  - It incorporates a cross-over design in which the patient will experience one intervention, followed by a washout period, and then experiences the second intervention.
  - The patient and clinician keep records of responses under both conditions to determine which is preferable. In this way, this design becomes a decision-making tool.
  - Several N-of-1 trials can be combined in a form of meta-analysis to show the effectiveness of an intervention.
- SSDs can be evaluated using visual techniques, documenting a change in the trend (slope) and level (magnitude) of the response over time.
- The split middle line is a line that is drawn through the median points in the baseline phase and extended into the intervention phase.
  - Also called a celeration line, it can be used a way of showing that the trend in response during the baseline phase does not continue into the intervention phase, thereby documenting a meaningful change.
  - A change from A to B is evaluated using the binomial test, which determined if there is a significant difference in the proportion of points that fall above and below the extended line.
- The two-standard-deviation band method involves calculating the mean and standard deviation of points in the baseline phase.
  - Lines are extended from baseline to intervention phase at two standard deviations above and below the mean.
  - If at least two consecutive data points in the intervention phase fall outside the two standard deviation band, changes from baseline to interventions are considered significant.
- Statistical process control (SPC) is a method of examining variability in data across baseline and intervention phases.
  - Upper and lower control limits are plotted at 3 standard deviations above and below the baseline mean. Data that fall outside these limits are considered too variable to be considered similar to baseline.
- Effect size can also be used to evaluate changes from baseline to intervention phases.
  - A commonly used method is evaluation of *nonoverlapping scores* from baseline to intervention. The percent of non-overlap is a reflection of the treatment effect.
  - When all intervention points overlap with baseline points, there is 0% nonoverlap, and the distribution of scores in both phases is conceptually superimposed, indicating no treatment effect.
- Generalization of single-subject out-comes is important as a measure of

external validity.

- Generalization can be provided by direct replication by repeating single-subject experiments across several subjects.
- Systematic replication demonstrates consistency of findings under conditions that are different from the initial study.
- Clinical replication involves replication of effects in typical clinical situations where treatment may be used in various combination to fit the patient's needs.
- Social validity has been used to indicate the application of single-subject study findings in real world settings.

CHAPTER 19

Exploratory Research: Observational Designs

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Distinguish between observational and experimental research.
2. Describe criteria for evaluating causality in observational research.
3. Distinguish between longitudinal and cross-sectional study designs.
4. Distinguish between prospective and retrospective study designs.
5. Describe the structure and design elements of a cohort study.
6. Describe the structure and design elements of a case-control study.
7. Identify strategies for minimizing bias or confounding in cohort and case-control studies.

■ Key Terms

Observational studies
Exploratory research
Descriptive research
Analytic research
Exposure
Risk factor
Dose-response relationship
Longitudinal studies
Prospective studies
Retrospective studies
Cross-sectional studies
Reverse causation
Cohort studies
Secondary analysis
Misclassification
Case-control study
Population-based study
Hospital-based study
Selection bias
Observation bias
Interviewer bias
Recall bias
Confounding
Matching
Propensity score matching

■ Chapter Summary

- Observational research can be classified as descriptive or analytic. *Descriptive research* characterizes populations by examining the distribution of health-related factors. *Analytic research* focuses on examining group differences to demonstrate how exposures can help explain observed differences.
- Observational research differs from experimental research in that there is no manipulation of variables and no assignment of

subjects to groups. Existing groups are identified by their shared history or health status. Observational designs are needed when it is not feasible to use experimental designs for demonstrating treatment effectiveness.

- A common use of observational designs is to draw causal inferences about the effects of a hypothesized risk factor. Risk factors can refer to factors related to the occurrence or prevention of a disorder, which may include personal characteristics, behaviors, or environmental exposures.
- To support a causal hypothesis, observational designs must be able to demonstrate several principles:
  - *Temporal sequence* establishes that a causative exposure precedes the outcome.
  - There should be a substantial association between the exposure and outcome, based on a statistical measure of risk.
  - The relationship should be based on a plausible *biological mechanism*.
  - Findings should demonstrate *consistency* across many studies using different samples under different conditions.
  - A dose-response relationship indicates that the outcome is related to increasing levels of exposure. Nonlinear relationships are also possible.
- Observational studies may be longitudinal, where researchers follow subjects over time, or cross-sectional, when groups of subjects are evaluated at one point in time. Longitudinal designs may be prospective, involving direct recording of data to define exposed and unexposed groups, or retrospective, utilizing exposure data that were collected in the past.
  - A major challenge for cross-sectional studies is establishing the correct temporal sequence of cause and effect. Reverse causation can occur when the specified "outcome" actually causes the "exposure".
  - A disadvantage of the retrospective study is the lack of investigator control of how past data were collected.
- A cohort study is a longitudinal investigation in which the researcher identifies a group of subjects who do not yet have the outcome. Exposed and unexposed cohort members are monitored for a period of time to ascertain and compare the incidence of new outcomes.
  - An *inception cohort* is a group that is assembled early in the development of a disorder. Exposed and unexposed persons are then followed to determine the longitudinal association of exposure with progression or remission of the disorder.
  - Cohort studies are inappropriate for studying rare or slowly developing outcomes because large numbers of subjects would need to be followed over long periods to detect a sufficient number of cases for analysis.
  - Subjects for cohort studies can be recruited from the general population or from special populations in which exposure to a risk factor is common. Subjects who are exposed should be as similar as possible to the unexposed group in all factors related to the outcome. No subjects should be immune to the outcome.
  - Major challenges for cohort studies include the potential for attrition and the possibility of misclassifying a subject's exposure status.
- A case-control study includes groups of individuals who are purposely selected on the basis of whether they have the health condition under study. *Cases* are those who have the target condition, while *controls* do not. The investigator then determines if these two groups differ with respect to their exposure histories.
  - A case definition refers to the diagnostic and clinical criteria used to identify someone as a case.

- A population-based study obtains cases and controls from the general population. In a hospital-based study, subjects are patients in a medical institution.
- Cases and controls must be chosen regardless of their exposure history, protecting the study against selection bias.
- Observation bias occurs when there is a systematic difference in the way information about the exposure is obtained from the study groups. Recall bias occurs when cases who have experienced a particular disorder remember their exposure history differently than do controls.
- Confounding occurs when extraneous variables that alter the risk of developing the disorder are unequally distributed across exposed and unexposed groups.

CHAPTER
20

Descriptive Research

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the purpose of descriptive research and how descriptive data can be used to generate hypotheses.
2. Describe the purposes of developmental and normative research.
3. Discuss the advantages and disadvantages of longitudinal and cross-sectional approaches for collecting descriptive data.
4. Discuss the structure and purpose of case reports.
5. Describe the impact of epidemiologic care reports.
6. Describe the function of historical research and the type of data used to synthesize historical data.

■ Key Terms

Descriptive research
Developmental research
Longitudinal studies
Natural history
Cross-sectional studies
Cohort effects
Period effects
Normative research
Case reports
Case series
Historical research
External criticism
Internal criticism

■ Chapter Summary

- Descriptive research is an observational approach designed to document traits, behaviors and conditions of individuals, groups, and populations.
- Descriptive studies can use quantitative or qualitative methods.
- Descriptive research is generally structured around guiding questions rather than hypotheses, but descriptive data will often lead to generation of hypotheses or theoretical propositions that can be tested using exploratory or explanatory techniques.
- Developmental research involves the description of developmental change and the sequencing of behaviors over time. Important developmental studies have provided the foundation for understanding human development in children and adults.

- Longitudinal developmental studies involve collecting data over an extended period of time to document behaviors as they naturally change.
- A cross-sectional study involves collection of data at one point in time to describe differences among members of a group, differentiating those who are at different developmental levels. This is often called a "snapshot" approach and is considered a more pragmatic approach for collecting data on large samples.
- Potential disadvantages of the cross-sectional approach include cohort effects, whereby individuals can vary in their developmental trajectory because of when they were born or what specific events occurred at certain points in time.

- Normative studies describe typical or standard values for characteristics of a given population. They can be directed toward a specific age group, gender, occupation, culture or disability. Norms are usually expressed as averages within a range of acceptable values.
  - Normal values can be generated for physiological measures, such as laboratory tests, or they can be generated as standardized scores that allow interpretation of responses relative to a normed value for a given group, such as IQ scores.
  - Studies that seek to establish normal values must be large to represent the full range of responses and may need to be established for subgroups.
  - Normal values are important references for determining when an individual needs intervention and when the person has reached an acceptable performance standard.
- Descriptive surveys are used to provide an overall picture of a group's characteristics, attitudes or behaviors.
  - Survey data may be used to establish risk factors for disease or dysfunction, to provide data for development of hypotheses, and to establish population characteristics that can be used for policy and resource decisions.
- Case reports provide a detailed account of an individual's condition or response to treatment but may also focus on a group, institution, or other social unit, such as a school or community.
  - The purpose of a case report is to reflect on practice, including diagnostic methods, patient management, ethical issues, innovative interventions, or the natural history of a condition.
  - Case reports are not considered a true form of research because they are not based on a rigorous or systematic methodology. However, they can be important contributions to evidence-based practice as they present new information that can be considered by others.
  - Case reports can be prospective, following a case as it progresses, or retrospective, reporting the results of case management after it is completed.
  - A *case series* is an expansion of a case report involving observation of several cases that have similarities.
  - Many journals publish case reports and will require specific formats.
  - Clinicians must take precautions to protect patient privacy in case reports and should obtain informed consent from patients.
  - *Epidemiologic case reports* describe one or more individuals, documenting unique or unusual health concerns. These types of reports are important contributions that will often lead to investigation of specific exposures that may cause the observed conditions.
- Historical research involves the critical review of events, documents, literature, and other sources of data to reconstruct the past in an

effort to understand how and why past events occurred and how that understanding can influence current practice or outcomes.

- Historical studies are intended to incorporate judgments, analyses, and inferences on the part of the researcher through interpretation of historic facts and observed relationships.
- Historical data are not just a collection of past events but a synthesis of data from the past that can be applied to understanding current conditions.

CHAPTER

21

Qualitative Research Methods

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Understand key differences between qualitative and quantitative methods.
2. Differentiate among common types of qualitative research approaches.
3. Understand how and why qualitative sampling differs from quantitative sampling.
4. Differentiate among the three common types of qualitative interviews.
5. Differentiate among common modes of qualitative analysis.
6. Understand how researchers maintain rigor/trustworthiness in qualitative research.
7. Read a qualitative article and evaluate the trustworthiness of the findings.
8. Understand how mixed methods research can be used to combine both qualitative and quantitative methods.
9. Choose a research article that uses a quantitative methodology and discuss how qualitative study can complement that study's purpose.

■ Key Terms

Qualitative research
Logical positivism
Naturalistic inquiry
Ethnography
Informants
Grounded theory
Constant comparison
Saturation
Phenomenology
Relationality
Temporality
Case studies
Field notes
Participant observation
In-depth interview
Unstructured interview
Interview guide
Structured interview
Semi-structured interview
Focus group interviews
Theoretical sampling
Coding
Content analysis
Trustworthiness
Credibility
Triangulation

Negative case analysis
Member checking
Transferability
Thick description
Dependability
Audit trail
Confirmability
Reflexivity
Mixed methods
Multimethod studies
Convergent designs
Sequential designs
Embedded designs
Multiphase designs

Chapter Summary

- Qualitative research is based on the belief that all interactions are inherently social phenomena, requiring a deep understanding of personal experience to explore behaviors and attitudes. To gain these insights, qualitative researchers conduct inquiry in natural settings.
- Qualitative data involve the collection of narrative information. This research is generally considered an inductive process, with conclusions being drawn directly from the data.
- In the systematic study of social phenomena, qualitative research serves three important purposes: describing groups of people and social-cultural contexts, generating hypotheses that can be tested by further research, and developing or expanding theory to explain observed phenomena.
- Qualitative research can be complementary to quantitative study, allowing one to inform the other regarding relevant variables and questions.
- Qualitative research addresses essential components of evidence-based practice with regard to sources of information, including patient preferences and providers' clinical judgments. These factors are influenced by understanding how patients and care providers view interactions with each other, family members, and colleagues within their environment.
- Qualitative research questions usually start out broad, seeking to understand the dynamics of how an experience influences subsequent behaviors or decisions, including why something happens. Questions can be added to the inquiry as data are collected that uncover new information. Qualitative studies do not usually specify conventional hypotheses.
- Although the concept of causation is not directly applicable to qualitative study, findings can offer interpretations about possible explanatory factors based on understanding an individual's personal responses.
- The most commonly applied traditions in qualitative study include ethnography, grounded theory, and phenomenology.
- Ethnography has its foundations in anthropology and has been defined by fieldwork where investigators immerse themselves in the settings and activities studied. Study participants become *informants* because they inform and teach the researcher about their lives and communities.
- Grounded theory is a formalized process of simultaneous data collection and theory development. Through a process of constant comparison, each new piece of information is compared to data already collected to determine agreements or conflicts. This iterative process continues until additional data yield no further insights, called

saturation. The researcher puts pieces of information in context, eventually leading to a theoretical premise. The derived theory is then formulated, based on factors that interact to explain behaviors.

- Phenomenology is the study of life experiences and the meanings individuals attribute to those experiences. Observation and interviews are used to reveal participants' ideas and thoughts related to their experiences within their lived environment (*spatiality*), regarding the influence of interactions and emotions (*relationality*), and perception of time (*temporality*).
- Case study research is used to understand an individual's behaviors, an institutional culture, or systems within natural settings. Through in-depth interviews, questionnaires, and review of documents, qualitative researchers draw conclusions that help to explain interactions and choices.
- Qualitative data collection includes several strategies, including observation, interviews, and focus groups.
  - Observation provides data about how social processes occur in real life situations, without external influences. *Field notes* are used to record descriptions of what is seen and heard. The observer steps back and monitors behaviors with no interaction.
  - In participant observation, often used in ethnographic research, the researcher becomes part of the community being studied. The intent is that data will be more informed because the researcher experiences are the same as those of the participants.
  - Interviews are usually conducted face to face. The purpose of an in-depth interview is to probe ideas and obtain the most detailed information possible. *Unstructured interviews* are open ended based on a list of general questions to stimulate discussion. In a *structured interview*, questions and response choices are predetermined, similar to using a questionnaire. A *semi-structured interview* uses a combination of fixed responses and open-ended questions, allowing some consistency in the interview content, but also allowing participants to contribute their own insights.
- Qualitative researchers use different approaches to non-probability sampling, including convenience, purposive, and snowball sampling. It is important to choose participants who represent a range of characteristics, and who can be a rich source of information.
- Determining an adequate sample size for qualitative study is a matter of judgment, often based on evaluation of the quality of information collected and the purpose of the research. Although often small, qualitative studies can incorporate larger samples depending on the environment and the scope of the question.
- Theoretical sampling is a process whereby participants are recruited and interviewed, and data are analyzed to identify initial concepts. New participants are brought in who are likely to expand on those concepts, until components of theory emerge, and data have reached the point of saturation.
- In the analysis of qualitative data, researchers will usually develop a coding scheme to identify themes or concepts that emerge from the data. Through a process of content analysis, inferences are drawn by interpreting the textual material.
- The quality of a qualitative study is described as its *trustworthiness* based on four criteria. C*redibility* refers to the degree of confidence in the truth of the findings. *Dependability* refers to how stable the data are over a span of time relevant to the study and the degree to which the study could be repeated with

similar results. *Confirmability* refers to ensuring that findings are due to the experiences and ideas of participants, rather than characteristics or preferences of the researcher. *Transferability* refers to the ability to apply findings to other people in similar situations, or an assessment of generalizability.

- Several methods are used to establish credibility of findings, including triangulation, which involves comparing various sources of data to confirm findings; negative case analysis, which involves discussing elements of data that do not support explanations emerging from the data; and member checking, which involves bringing participants or other stakeholders into the process to validate findings and offer feedback.
- Mixed methods research involves combination of qualitative and quantitative methods. Designs incorporate elements of data collection from both traditions and integrate the findings to explain the phenomenon under study.

CHAPTER

22 Descriptive Statistics

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe methods for graphic presentation of distributions.
2. Calculate measures of central tendency and discuss their appropriate applications.
3. Define measures of variability including range, percentiles, variance, and standard deviation.
4. Discuss the concept of sum of squares and its meaning in defining variance.
5. Explain the properties of the normal curve.
6. Explain the purpose of standardized scores in relation to the normal curve and standard deviation.

■ Key Terms

Distribution
Frequency distribution
Grouped frequency distribution
Histogram
Line plot (frequency polygon)
Stem-and-leaf plot
Exploratory data analysis (EDA)
Normal distribution
Skewed distribution
Central tendency
Mode
Median
Mean
Range
Percentiles
Quartiles
Interquartile range
Box plot
Deviation scores
Sum of square (SS)
Mean square (MS)
Variance (s^2)
Skewness
Standard deviation (s)
Coefficient of variation (CV)
Standardized scores (z-scores)

Chapter Summary

- Descriptive statistics are used to characterize the shape, central tendency, and variability within a set of data, called a distribution.
- Frequency distributions provide an ordered list of scores within a distribution and the number of times that score occurs. When continuous data are used, frequency distributions can contain data grouped into meaningful intervals.
- Histograms, line graphs, and stem-and-leaf plots can be used to graphically display distributions.
- The shape of data can demonstrate how scores are distributed. The normal curve is a symmetrical bell-shaped curve. A skewed distribution is asymmetrical, with fewer scores at one end forming a "tail." A *positively skewed* distribution has a longer tail on the right, and a *negatively skewed* distribution has a longer tail on the left.
- Three measures of central tendency, or *averages*, can be used to characterize data. The mode is the score that occurs most frequently in a distribution. The median is that value above which there are as many scores as below it, dividing a rank ordered distribution into two equal halves. The mean of the distribution is the sum of a set of scores divided by the number of scores, the value most people call the "average."
- Variability can be expressed in several ways. The range is the difference between the highest and lowest scores. Percentiles divide data into 100 equal portions. If a score is in the 20th percentile, that score is higher than 20% of the scores in that distribution. Quartiles divide a distribution into four equal parts. Q_1, Q_2, and Q_3 correspond to 25%, 50% (the median), and 75% of the distribution. The interquartile range is the range of scores within the middle 50% of the distribution, between Q_1 and Q_3. These values are often graphed as box plots which show the median, the interquartile range, and the minimum and maximum scores.
- Variability is a measure of the spread of scores within a distribution. As variability increases, scores are more spread out around the mean. It is assessed by taking the sum of deviation scores, or difference of each score from the mean. These values are squared to ignore minus signs, and their sum indicates the degree of variance within the distribution. These values are, therefore, referred to as sum of squares (SS).
- The sum of squares represents the full variance in the distribution. However, because values were squared, the square root of this value is used as a measure of variability within a distribution, called the standard deviation (SD).
- In a normal distribution, proportions of the curve are standardized by converting standard deviation units to a *z* score. Approximately 68% of the curve represents the distance from –1 to +1 standard deviation units, or $z = \pm 1.0$. Approximately 95% of the curve falls between ±2 SD, and approximately 99% falls within ±3 SD.
- Using *z* scores, we can determine the area under the normal curve that is bounded by any two values.

CHAPTER 23

Foundations of Statistical Inference

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the concept of probability in relation to observed outcomes.
2. Describe the concept of sampling error.
3. Define and interpret a confidence interval for the mean.
4. Distinguish between research and null hypotheses.
5. Explain the difference between the null hypothesis and the alternative hypothesis.
6. Define Type I and Type II errors.
7. Discuss the purpose of a priori and post hoc statistical power analysis.
8. Discuss the determinants of statistical power.
9. Explain the difference between a one-tailed and two-tailed test.
10. Explain the conceptual difference between parametric and nonparametric statistics.

■ Key Terms

Inferential statistics
Probability (p)
Sampling error of the mean
Standard error of the mean ($s_{\bar{X}}$)
Point estimate
Interval estimate
Confidence interval (CI)
Null hypothesis (H_0)
Alternative hypothesis (H_1)
Superiority trial
Non-inferiority trial
Not significantly different
Type I error
Type II error
Level of significance
Alpha (α)
Beta (β)
Power
Effect size (ES)
Effect size index
Conventional effect sizes
Statistical significance
Clinical significance
Publication bias

z-ratio
Critical region
Critical value
Two-tailed test
One-tailed test
Parametric statistics
Homogeneity of variance
Nonparametric tests
Kolmogorov-Smirnov test of normality
Shapiro-Wilk test of normality

■ Chapter Summary

- Inferential statistics involve a decision-making process that allows us to estimate unknown population characteristics from sample data.
- Assumptions for inferential statistics are based on two concepts of statistical reasoning: probability and sampling error.
  - Probability is the likelihood that any one event will occur, given all possible outcomes. A lowercase *p* is used to represent probability as a decimal, from 0 (no probability) to 1.00 (absolute probability). If $p = .04$, it means there is a 4% probability that an event will occur.
  - Sampling error refers to the difference between true population values and sample values. Because we rarely know population values, we make estimates based on sample data using standard error**.** For means, the *standard error of the mean* is used to estimate the population mean, based on the ratio of the sample standard deviation and the $\sqrt{n}$. As *n* increases, the standard error will decrease, indicating that the estimated values should be closer to the population mean.
- Any single group statistic, such as a mean, is considered a *point estimate*. However, to better estimate population values, it is more helpful to determine an interval that is likely to contain the population value. This is called a confidence interval (CI)**.** If we calculate a 95% CI, it means that we are 95% sure that the true population value will fall within that range.
- The null hypothesis (H_0) is a statistical hypothesis of no difference, which will be tested by statistical procedures. The alternative hypothesis (H_1) is based on the research hypothesis, predicts an effect, and can be expressed in a directional or non-directional format.
  - The purpose of statistical testing is to try to disprove H_0. Because we cannot prove a negative, we can only legitimately *reject* H_0 if an effect is demonstrated or *not reject it* if we cannot demonstrate an effect.
  - Because we deal with sample data, when we make a decision to reject or not reject H_0, we do so knowing we may be making a correct or incorrect decision. A Type I error occurs when we conclude that there is a significant effect when there is none, that is, we incorrectly reject H_0 when it is true. A Type II error occurs when we find no effect when there really is one, that is, we incorrectly reject H_0 when it is false.
  - When we specify a null hypothesis for a non-inferiority trial, we predict that the experimental treatment will be worse than the standard treatment by a certain amount, called the non-inferiority margin**.** The alternative hypothesis will be that the experimental treatment will not be worse than the standard treatment.
- We specify a level of significance, denoted as alpha (α), that indicates a threshold we will use for committing a Type I error. Therefore, if α is set at .05, it means we will reject H_0 only if we find a significant difference with ≤5% chance of being incorrect. Although this is an arbitrary threshold, it has become accepted as standard.

- When a statistical test is performed to determine if a true effect has occurred, such as a difference between group means, the test will result in a *p* value that indicates the likelihood of making a Type I error if we decide the groups are different. Using .05 as the standard threshold, for example, $p = .04$ would be considered significant.

- The probability of making a Type II error is **beta (β)**, which is the probability of failing to reject a false null hypothesis. If $\beta = .20$, then there is a 20% chance that we will not reject H_0 when it false.

- The complement of β error, $1 - \beta$, is considered the power of a test. A more powerful test is more likely to find a significant difference if it exists.
  - Power involves four interdependent concepts (PANE): power ($1 - \beta$), alpha level of significance (α), sample size (n), and effect size. With a larger sample, a test will be more powerful.
  - The effect size (ES) is a measure of the degree to which H_0 is false, or the size of the effect of the independent variable. An effect size index is a unitless standardized value that allows comparisons of effect size across samples and studies. There are several types of ES indices, each with conventional effect sizes that are considered small, medium, and large.
  - *A priori* power analysis can be used to estimate sample size during planning of study by estimating the expected effect size, desired power, and levels of α and β. *Post hoc* analysis can be used to estimate achieved power when a study is completed and a significant effect is not found.

- In all statistical testing, it is important to consider the difference between statistical significance based on probability and clinical significance based on expertise, experience, and an understanding of theoretical mechanisms that puts statistical results in context. All research results should be considered in terms of whether outcomes are meaningful. A statistically significant effect doesn't mean it is important, and vice versa.

- In statistical testing, the size of the effect is based on a standardized score, the *z* score. Based on proportions of the normal curve, we can determine if a *z* score falls within tails of the curve that represent 5%. Anything that falls within the upper or lower 5% is considered an unlikely event and therefore would represent a significant effect.
  - A two-tailed test is used when the alternative hypothesis is expressed without direction, allowing for the possibility that the effect could be positive or negative. A one-tailed test is used when the hypothesis is directional, indicating that a difference is expected only in one direction.

- Parametric statistics are used for statistical testing based on three assumptions.
  - Samples are randomly drawn from a parent population with a normal distribution.
  - Variances in the samples being compared are roughly equal, or *homogeneous*.
  - Data are measured on the interval or ratio scales.

- Nonparametric statistics are used when these assumptions cannot be made. These statistical procedures are based on ranked data and are useful with ordinal or nominal data.

CHAPTER

24

Comparison of Two Means: *t-Test*

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the components of the statistical ratio for examining the difference between two means.
2. Discuss the assumptions for using statistical tests.
3. Discuss the application of the *t*-test for paired and unpaired research designs.
4. Describe the use of confidence intervals in statistical testing.
5. Interpret computer output for the paired and unpaired *t*-test.
6. Determine power and sample size estimates for studies comparing two means.
7. Explain why the use of multiple *t*-tests is inappropriate for statistical decision making.

■ Key Terms

t-test
Error variance
Statistical hypothesis (H_0)
Unpaired *t*-test (independent)
Homogeneity of variance
Standard error of the difference between means
Critical value
Degrees of freedom (*df*)
Levene's test for equality of variances
Two-tailed test
One-tailed test
Confidence intervals
Paired *t*-test
Standard error of the difference scores
Standardized mean difference (SMD)
Effect size

■ Chapter Summary

- The *t*-test is used to compare two means, either from two independent groups or from two repeated measures.
- The statistic is based on the ratio of variance *between groups* divided by the variance *within groups*. The denominator of the *t*-test is called the standard error of the difference between means.
- An assumption for the *t*-test, like other parametric statistics, is that there is equality of variances among the two groups, or

homogeneity of variance. When variances are significantly different, the *t*-test has to be adjusted. Levene's test is used to compare the variances of the two groups.

- Each group has *n*-1 degrees of freedom.
- The calculated value of *t* is compared with a critical value that is determined by degrees of freedom, level of significance, and one- or two-tailed tests.
- Confidence intervals can be determined for the difference between means. These values present important information to judge the true nature of the difference between means.
- The *independent* or unpaired *t*-test is used when two independent groups are compared, usually formed by random assignment. The computer output for the unpaired *t*-test automatically generates two lines of data, one for *equal variances assumed* and another for *equal variances not assumed.* The researcher must decide which data to use based on the test for homogeneity of variance.
- The paired *t*-test is used when the two levels of the independent variable represent a repeated measure. Because the number of measures must be equal across the two treatment conditions, it is unnecessary to test for homogeneity of variance with the paired test.
- Power of the *t*-test is estimated using the effect size index *d*, which expresses the difference between two sample means in standard deviation units. Conventional effect sizes are small: $d = .20$, medium: $d = .50$, large: $d = .80$.
- Statistical reports for the *t*-test should include the means (±SD), the mean difference, confidence intervals, one- or two-tailed test, the calculated value of *t*, degrees of freedom, value of *p*, and effect size.
- It is inappropriate to use the *t*-test for comparing more than two means, as this inflates the Type I error associated with the comparisons. The analysis of variance is recommended when more than two means are compared.

CHAPTER 25

Comparison of More than Two Means: Analysis of Variance

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the application of the analysis of variance for one-way and two-way designs.
2. Discuss the difference between main effects and interaction effects in a two-way design.
3. Discuss the difference between between-subjects and within-subjects analyses.
4. Interpret computer output for analysis of variance.
5. Interpret power and effect size for analysis of variance.

■ Key Terms

Analysis of variance (ANOVA)
One-way ANOVA
Partitioning the variance
Between-groups variance
Within-groups (error) variance
Homogeneity of variance
Sum of squares (SS)
Between-groups sum of squares (SS_b)
Error sum of squares (within groups) (SS_e)
Total sum of squares (SS_t)
Degrees of freedom
Mean square (MS)
Between-groups mean square (MS_b)
Error mean square (within groups) (MS_e)
F-ratio
Multiple comparison
Eta squared (η^2)
Cohen's *f*
Omega squared (ω^2)
Two-way ANOVA
Factorial design
Main effects
Marginal means
Interaction effects
Repeated measures ANOVA (within-subjects design)
Sphericity
Mauchley's test of sphericity
Greenhouse-Geiser correction
Huynh-Feldt correction
Mixed analysis of variance

Chapter Summary

- The analysis of variance (ANOVA) is used to compare means for more than two groups or repeated measures.
- The ANOVA is a parametric test based on assumptions of homogeneity of variance, normal distribution, and interval/ratio data.
- The one-way ANOVA is used with one independent variable or factor, with three or more levels.
- Total variance in a set of data reflects two sources of variance: a treatment effects (between groups) and unexplained error (within groups).
- Variance is calculated as the sum of squares (*SS*). A separate *SS* will be calculated for between groups, within groups, and total variance.
- Levene's test is used to determine homogeneity of variance. If the test is not significant, variances are assumed to be equal.
- The format for a table reporting the results of an ANOVA includes sum of squares, mean square (*MS*), and degrees of freedom (*df*) for each source of variance, and the calculated F ratio and probability for each comparison. For between groups variance $df = k - 1$, for within groups variance $df = n - k$, and total $df = N - 1$.
- The F ratio is calculated as the MS_b/MS_e. Subscripts b = between groups, e = error term.
- The ANOVA can include values for effect size and power. Effect size indices for the ANOVA are eta **squared** (η^2) or partial eta squared (η_p^2) and Cohen's f. These are calculated using values for the SS_b and SS_e. Conventional effect sizes have been proposed for these indices.
- The two-way ANOVA is used with a factorial design, when two independent variables are used.
- Variance in a two-way design includes main effects for each independent variable as well as the interaction of the two variables. When interaction effects are significant, main effects are usually ignored.
- The repeated measures ANOVA is used when all subjects are tested under all levels of the independent variable.
- Homogeneity of variance for repeated measures is based on the assumption of sphericity. This is tested using Mauchley's test of sphericity.
- Repeated measures can be used in a two-way design, or in a mixed design with one repeated measure and one independent measure.

CHAPTER

26

Multiple Comparisons

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Explain the difference between per comparison and familywise error rate.
2. Define the minimal significant difference (MSD).
3. Calculate the MSD for several multiple comparison tests.
4. Discuss the purpose of trend analysis.
5. Interpret computer output for multiple comparisons tests.

■ Key Terms

Multiple comparisons tests
Per comparison error rate (α_{PC})
Familywise error rate (α_{FW})
Type I error rate
Bonferroni correction
Post hoc multiple comparisons
Minimum significant difference (MSD)
Tukey's honestly significant difference (HSD)
Studentized range (q)
Student-Newman-Keuls test (SNK)
Scheffé comparison
Simple effects
Planned comparisons
Complex contrasts
Trend analysis

■ Chapter Summary

- Multiple comparisons are used to explore differences between means following an analysis of variance.
- Multiple comparisons provide adjustments for the level of significance to account for the use of several comparisons. More liberal procedures will have greater power but also a greater chance of committing a Type I error. More conservative approaches will have lower power but decrease the probability of Type I error.
- The per comparison error rate (α_{PC}) is the probability associated with a single comparison. The familywise error rate (α_{FW}) represents the cumulative probability of making at least one Type I error in a set of statistical comparisons.
- The Bonferroni correction is an adjustment determined by dividing the level of

significance (e.g. .05) by the number of comparisons, correcting the probability that is needed to achieve significance.

- *Post hoc* multiple comparisons are performed after an ANOVA and may involve comparison of all pairwise differences.
- The most commonly used *post hoc* tests include Tukey's honestly significant difference (HSD), the Student-Newman-Keuls (SNK) comparison, and Scheffé's comparison. The SNK procedure is the most liberal of the three, Scheffé is most conservative, and Tukey provides a good balance between Type I and Type II error.
- The *post hoc* tests are based on comparison of a difference between means with a minimum significant difference (MSD). The mean difference must be equal to or greater than the MSD to be considered significant. The Tukey test uses one MSD for all pairwise comparisons. The SNK procedure uses different MSD depending on how far apart the means are. The Scheffé test is based on the *F* statistic to determine the MSD.
- *Post hoc* tests can be run for one-way or two-way designs and for repeated measures. In two-way designs, comparisons may include marginal means for main effects or means for interaction.
- *A priori* multiple comparisons are planned comparisons, decided before a study is run, focusing on specific comparisons rather than all possible pairwise comparisons.
- Trend analyses are used to look at patterns of a responses over a continuous variable such as age or time. Trends are generally described as linear or nonlinear. The most frequently encountered nonlinear trend is quadratic, which means scores vary in direction or rate of change. Trends can be tested using an ANOVA that can partition variance into trend components.

CHAPTER 27

Nonparametric Statistics for Group Comparisons

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the role of nonparametric statistics and reasons for their use.
2. Compare parametric and nonparametric tests for specific applications.
3. Explain the procedure for ranking a distribution of scores with and without ties.
4. Calculate and interpret results of the Mann-Whitney *U* test, the Kruskal-Wallis ANOVA, the sign test, the Wilcoxon signed-ranks test, and the Friedman ANOVA.
5. Interpret multiple comparison tests for the Kruskal-Wallis and Friedman analyses of variance.

■ Key Terms

Nonparametric statistics
Mann-Whitney *U* test
Kruskal-Wallis analysis of variance by ranks (*H*)
Sign test
Binomial test (*x*)
Wilcoxon signed-ranks test (*T*)
Friedman two-way analysis variance by ranks (χ_r^2)

■ Chapter Summary

- **Nonparametric tests** are statistical procedures that do not make the same assumptions of data as parametric statistics. They are often referred to as *distribution-free tests.*
  - Data used in nonparametric statistics are reduced to ranks, and comparisons are based on median distributions, not means.
  - The two major criteria for choosing a nonparametric test are that data are at the ordinal or nominal level of measurement or that assumptions of normality and homogeneity of variance cannot be satisfied.
  - The Kolmogorov-Smirnov and Shapiro-Wilk tests can be used to determine if data follow a normal distribution.
- Nonparametric tests can be compared to parametric procedures for *power efficiency*, which is a test's relative ability to identify significant differences for a given sample size. A rule of thumb to determine sample size is to add 15% to the required sample size if the data were used with a parametric procedure.

- To rank scores in a distribution, data are listed from smallest to largest, taking into account negative signs. If two or more scores are tied, each is given the same rank, which is the average of the ranks they occupy.
- The **Mann-Whitney *U* test** is used to test the difference between independent samples and is analogous to the unpaired *t*-test.
  - Scores are ranked for the total sample, and the sums of ranks within each group are then determined. Under the null hypothesis, we would expect the groups to be equally distributed with regard to high and low ranks, and the mean rank for each group would be the same.
  - The effect size index for *U* is a correlation coefficient based on *z*.
- The **Kruskal-Wallis One-Way Analysis of Variance by Ranks** is the nonparametric analog of the one-way analysis of variance (ANOVA), used with more than two independent groups.
  - Scores are ranked for the total sample, and the sums of ranks within each group are then determined. Under the null hypothesis, ranks would be equally distributed across all groups.
  - The test statistic is *H*, which is based on the chi-square distribution (see Chapter 28).
  - *Dunn's multiple comparison* is a nonparametric test for pairwise comparisons for the Kruskal-Wallis test.
  - The effect size index for *H* is eta squared, $\boldsymbol{\eta^2}$.
- The **sign test** is used to compare paired scores, analogous to the paired *t*-test. It is based on the binomial test which considers dichotomous data.
  - The test looks at each pair of scores and determines the direction of difference. Under the null hypothesis, the distribution of + or – scores under each condition will not be different.
  - The test statistic for the binomial test is *x*.
- The **Wilcoxon signed-ranks test** is a more powerful comparison for two related groups, also analogous to the paired *t*-test.
  - Difference scores for each pair are ranked. Under the null hypothesis, there would be an equal representation of + or – scores among higher and lower ranks.
  - The test statistic is *T*.
  - The effect size index for the Wilcoxon test is based on a correlation coefficient, similar to *U*.

See the Chapter 27 Supplement on Estimating Power and Sample Size for description of using G*Power for the Mann-Whitney U test and the Wilcoxon signed-ranks test.

- The **Friedman two-way analysis of variance by ranks** is the nonparametric equivalent of the one-way repeated measures ANOVA, used with three or more repeated conditions.
  - Scores are ranked within each subject across the repeated conditions, and the sums of ranks for each condition are determined. Under the null hypothesis, we would expect high and low ranks to be evenly distributed across all conditions.
  - The test statistic is $\boldsymbol{\chi_r^2}$, which is based on the chi-square distribution.
  - *Dunn's multiple comparison* is used as a nonparametric multiple comparison procedure.
  - The effect size index for the Friedman test is the Kendall coefficient of concordance, *W*.

CHAPTER 28

Measuring Association for Categorical Variables: Chi-Square

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the use of chi-square in research.
2. Discuss the purpose of a goodness-of-fit model.
3. Discuss how standardized residuals help to explain significant chi-square values.
4. Calculate chi-square for 2 × 2 tables and interpret results.
5. Discuss issues related to small samples sizes for interpretation of chi-square.
6. Describe other measures of association for nominal variables.

■ Key Terms

Chi-square statistic, χ^2
Goodness of fit
Uniform distribution
Known distribution
Standardized residuals
Contingency table
Expected frequencies
Yate's continuity correction
Fisher's exact test
Phi coefficient
Odds ratio
McNemar test

■ Chapter Summary

- Many research questions involve variables that are measured in categories, whether nominal or ordinal. Analysis of these variables involves proportions or frequencies within each category.
- Proportions are analyzed in terms of how they meet or depart from chance expectations. For example, a coin toss has two possible outcomes (heads or tails). If we toss a coin several times, assuming the coin is fair, we should see a 50:50 outcome in the long run. We can count the number of observed heads or tails to determine if it varies significantly from this expected distribution.
- Chi-square, χ^2, tests the difference between observed and expected frequencies.
  - The test is based on the assumption that frequencies represent individual counts, and categories are exhaustive and mutually exclusive. No one individual is represented in more than one category.

- Chi-square can be used to test goodness of fit of data with uniform or known distributions. The null hypothesis states that there is no difference between observed and expected counts.
  - A *uniform distribution* is one in which we would expect an equal proportion of individuals within each category.
  - In a *known distribution*, we can test observed counts against the known distribution of individuals within certain categories, often based on population data.
  - Critical values of χ^2 are based on *k-1* degrees of freedom, where *k* is the number of categories (see Appendix Table A-5). Calculated values must be equal to or greater than the critical value to be significant.
  - A significant test indicates that there is a difference between the observed frequencies and expected chance frequencies for each category. It does not mean that the number of counts within categories are different from each other.
  - Standardized residuals reflect the difference between expected and observed frequencies. They provide a basis for distinguishing which categories depart most significantly from the expected values.
- Chi-square is used most often as a measure of association between two categorical variables, with data organized within a contingency table or crosstabulation that crosses the two variables (with R rows and C columns).
  - The null hypothesis states that there is no association between the two variables, and that the expected frequencies across all categories will represent the same proportions as the observed frequencies.
  - Tests of significance for contingency tables are tested with (R-1)(C-1) degrees of freedom.
  - Yates' continuity correction can be used to adjust values when any cells within the contingency table have expected frequencies less than 5.0. This test is controversial, however, and is considered overly conservative.
  - Fisher's exact test is used instead of χ^2 when expected frequencies are less than 1 in some cells.
- The effect size index for χ^2 is given the notation *w* and is equal to $\sqrt{\frac{\chi^2}{N}}$.
- Other measures of effect are often calculated with χ^2.
  - The *phi coefficient*, ϕ, is interpreted as a correlation coefficient for 2×2 tables and is equivalent to *w*.
  - The *contingency coefficient*, *C*, is used when tables are larger than 2×2.
  - *Cramer's V* is used when tables are asymmetrical.
  - The odds ratio (OR) can also be used as an effect size measure with 2×2 tables (see Chapter 34).
- The McNemar test is a form of χ^2 used with 2×2 tables involving correlated samples, where subjects are their own controls. This is especially useful with pretest-posttest designs when the dependent variable is categorical.
  - The odds ratio can be used as an effect size index with the McNemar test, but it is adjusted to account for the matched values.

CHAPTER
29

Correlation

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the role of correlation for data analysis.
2. Describe the use of a scatterplot in correlation analysis.
3. Discuss the importance of distinguishing between linear and curvilinear relationships in correlation analysis.
4. Discuss the meaning of significance for a correlation coefficient.
5. Describe how variations of the Pearson *r* can be used to study relationships with continuous and dichotomous variables.
6. Discuss the importance of distinguishing between causation and correlation.
7. Discuss factors that can influence generalization of a correlation coefficient.
8. Describe the meaning of partial correlation for understanding the relationships among several variables.

■ Key Terms

Correlation
Scatter plot
Correlation coefficient
Pearson product-moment coefficient of correlation, *r*
Spearman rank correlation coefficient, r_s
Kendall's tau-b
Phi coefficient, ϕ
Point biserial correlation coefficient, r_{pb}
Rank biserial correlation coefficient, r_{rb}
Partial correlation
Zero order correlation
First-order partial correlation

■ Chapter Summary

- Correlation refers to the association between two variables, *X* and *Y*, reflecting the degree to which high scores in *X* are associated with higher scores in *Y*, and vice versa.
- A scatter plot provides a visual demonstration of that relationship. The closer points fall to a straight line, the higher the degree of correlation.

- The relationship between two variables can be positive, when high and low values are related, or negative, when high values on *X* are associated with lower values on *Y*.
- Correlation can be described quantitatively by a correlation coefficient, symbolized by *r*. Coefficients can usually range from -1.0 (a negative relationship) to +1.0 (a positive relationship).
  - The magnitude of the correlation coefficient indicates the strength of the association between *X* and *Y*. The closer to ±1.00, the stronger the relationship. A negative sign indicates the direction of the relationship only.
- Meaningful interpretation of correlation depends on certain assumptions:
  - Subjects' scores represent an underlying normal population
  - Each subject contributes an *X* and *Y* score
  - *X* and *Y* are independent (no one score is a component of the other)
  - The relationship between *X* and *Y* is linear
  - Values of *X* and *Y* are observed, not controlled or manipulated
- There is no consensus on what value of a correlation coefficient is considered strong or weak. Interpretations must be made within the context of what is being measured and how the correlation will be applied.
- The Pearson produce-moment coefficient of correlation is the most commonly applied correlation statistic for parametric data. The Pearson correlation is symbolized by *r*, but the Greek letter rho, ρ, is used to represent the population parameter.
  - The null hypothesis states that there is no relationship between two variables: H_0: $\rho = 0$.
  - Alternative hypotheses may be expressed with or without direction.
- The correlation coefficient can be tested for significance against a critical value with $n - 2$ *df*. The calculated value must be equal to or greater than the critical value to be significant.
  - Confidence intervals can also be constructed around the correlation coefficient.
- The effect size for correlation is *r*, which can be used to determine power and sample size.
- Two commonly used nonparametric correlation statistics are the Spearman rho, r_s, and the Kendall tau-b, τ. These statistics look at the common rankings of *X* and *Y*.
- Several statistics can be used to correlate dichotomies.
  - The phi coefficient, ϕ, is used when both *X* and *Y* are dichotomous.
  - The point biserial correlation coefficient, r_{pb}, is used to correlate one dichotomous variable with a continuous variable. It is equivalent to the independent *t*-test.
  - The rank biserial correlation, r_{rb}, is used to correlate one dichotomous and one ranked variable. It is equivalent to the Mann-Whitney *U* test.
- An important caveat in the interpretation of correlation is that it must be interpreted relative to its clinical significance, not simply its statistical significance.
- Correlation is a measure of *covariance* between two variables. It cannot be used to compare two variables or to establish a significant difference between two distributions.
- Correlation does not imply causation. The presence of an association between two variables may be due to a third variable that is correlated with both of them. Partial

correlation refers to the correlation of *X* and *Y* with the effect of a third variable removed. For instance, if *X* and *Y* are correlated, it is possible that a third variable, *Z*, is responsible for their relationship. If we look at the correlation of both *X* and *Y* with *Z*, then we can partition out the variance that is only explained by the relationship between *X* and *Y*. This would be the partial correlation with the effect of *Z* removed.

CHAPTER

30

Regression

Chapter Overview

Objectives

After completing this chapter, the learner will be able to:

1. Discuss the purpose of linear regression.
2. Explain the meaning of the line of best fit.
3. Explain how r^2 and standard error of the estimate help the researcher to determine the accuracy of the fit of a regression equation.
4. Describe the components of a multiple regression equation.
5. Describe the application of stepwise regression analysis.
6. Discuss how nonlinear regression can be used to better reflect some distributions.
7. Discuss the use of logistic regression for predicting dichotomous outcomes.
8. Discuss the application and interpretation of analysis of covariance.
9. Interpret computer output for regression procedures.

Key Terms

Coefficient of determination (r^2)
Simple linear regression
Independent variable (X)
Dependent variable (Y)
Regression line
Regression equation
Regression constant
Regression coefficient
Line of best fit
Residual
Adjusted R^2
Standard error of the estimate (SEE)
Analysis of variance of regression
Standardized residuals
Multiple regression
Beta weigh
Multicollinearity
Partial correlation
Tolerance level
Variance inflation factor (VIF)
Stepwise multiple regression
R^2 change
Forward inclusion
Backward deletion
Hierarchical regression
Dummy variables
Coding
Nonlinear regression
Quadratic curve
Polynomial regression
Logistic regression
Maximum likelihood estimation

Odds ratio (OR)
Indicator variable
Cox & Snell R square
Nagelkerke R square
Homer-Lemeshow test
Adjusted odds ratio
Propensity scores
Analysis of covariance (ANCOVA)
Covariate
Adjusted means
Homogeneity of slopes

Chapter Summary

- Regression is an extension of correlation specifically designed to analyze shared variance to establish how well one variable, or a set of variables, can predict and outcome.
- The square of the correlation coefficient, r^2, called the coefficient of determination, is a measure of accuracy of prediction. It indicates the proportion of variance in X that can be explained by knowing the variance in Y.
- Simple linear regression involves the development of an equation to calculate predicted values of Y from X. The variable designated X is the *independent variable*, and the variable designated Y is the *dependent* variable. Both X and Y variables are expected to be continuous.
- A scatterplot is the best way to visually understand the degree of relationship between X and Y. For the purposes of prediction, a line is drawn, called the line of best fit or the regression line, which "best" describes the orientation of data points.
- The algebraic representation of the regression line is given by $\hat{Y} = a + bX$, where $\hat{Y}$ (*Y-hat*) is the predicted value of Y.
  - The term a is called a regression constant. It is the Y-intercept, representing the value of Y when $X = 0$. This can be a positive or negative value.
  - The term b is the regression coefficient. It is the slope of the line, which is the rate of change in Y for each one-unit change in X.
  - The regression equation can be used to predict a value of Y by knowing the value of X. However, unless prediction is perfect, and all points fall directly on the regression line, this prediction may not be accurate for every person. The deviation of an individual score from the regression line (the predicted score) is called the residual, indicating the degree of error in the regression line.
- Prediction accuracy can be analyzed in several ways.
  - The value of R^2 indicates the proportion of variance in the dependent variable that is explained by the independent variable.
  - The adjusted R^2 is slightly smaller than R^2, representing a chance-corrected value.
  - The standard error of the estimate (SEE) reflects the variance of errors on either side of the regression line, or the residuals.
- Regression analysis tests the null hypothesis H_0: $b = 0$ and is analogous to testing the significance of the correlation between X and Y.
  - This hypothesis can be tested using an analysis of variance of regression which tests how much variance in the scores is explained by the regression line.
  - Correlation coefficients can be used to test the relationship between X and Y.

- The slope of the regression line can be tested using a *t*-test, indicating if the slope is significantly different from zero.

- There is an assumption that for any given value of *X*, a normal distribution of *Y* scores exists in the population. The observed value of *Y* at a particular value of *X* is one random value from this distribution.

- Multiple regression allows for prediction of *Y* using a set of several independent *X* variables. Outcome variables are often better explained by a series of variables, rather than just one.
  - The multiple regression equation accommodates several predictor variables: $\hat{Y} = a + b_1X_1 + b_2X_2 + ... b_kX_k$.
  - The regression coefficients (*b*) are essentially weights that indicate the contribution of each variable to the predicted outcome.
  - Standardized regression coefficients, called beta weights, are essentially *z* scores, allowing comparison of regression coefficients.

- Collinearity occurs when independent variables in a regression are correlated with each other and therefore provide redundant information.
  - Partial correlation can show how one variable is related to *Y*, accounting for the contributions of other independent variables.
  - Tolerance level and variance inflation factor (VIF) reflect the degree to which collinearity occurs. The lower the tolerance and the higher the VIF, the more collinearity is present.

- The effect size index for regression is R^2, which can also be converted to f^2 for use with G*Power.

- Stepwise multiple regression is a procedure whereby variables are entered into a regression equation one at a time based on the degree of explained variance. The process stops when no further variables add information to explain *Y*.

- Hierarchical regression involves entering sets of variables in sequential stages to examine relationships of subsets of variables.

- Although dependent and independent variables are expected to be continuous measures for use in regression, many variables are measured in categories. Dummy variables involve coding so that categorical variables can be included in the regression.

- Nonlinear regression, or polynomial regression, is used when the relationship between *X* and *Y* is not linear. Other than linear relationships, the most common trend is based on one turn, or a quadratic curve.

- Logistic regression is another form of prediction where dependent and independent variables are categorical, predicting group membership.
  - Variables are coded 0 and 1, with 1 typically representing the more adverse condition.
  - Odds ratios (OR) can be generated as part of logistic regression which indicate the odds of an outcome based on input variables. Adjusted odds ratios account for the influence of other variables in the equation in projecting an odds ratio for each independent variable.

- Propensity **scores** are outcome measures from logistic regression that reflect probabilities associated with the set of independent variables. These propensity scores can be used as a way of matching subjects in studies when matching has to account for several variables.

- The analysis of covariance (ANCOVA) is a form of ANOVA that is combined with regression to account for the influence of covariates in studying the relationship between independent and dependent variables. When groups differ on a relevant variable at baseline, the ANCOVA essentially adjusts scores so that the groups start off looking alike on those variables, reducing confounding that can be introduced when baseline characteristics affect group responses.

CHAPTER

31

Multivariate Statistics

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Explain the application and purposes of exploratory and confirmatory factor analysis.
2. Describe the purpose of structural equation modeling and its relationship to theory.
3. Describe the purpose of cluster analysis for describing subgroups in a set of individuals.
4. Describe the application of multivariate analysis of variance and discriminant analysis in the study of multiple dependent variables.
5. Describe the purpose of survival analysis and how survival rates and hazard ratios can be interpreted for different groups.

■ Key Terms

Exploratory factor analysis (EFA)
Principal components analysis (PCA)
Confirmatory factor analysis (CFA)
Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy
Bartlett's test of sphericity
Communality
Extraction
Principal axis factoring
Eigenvalue
Factor matrix
Rotated factor matrix
Varimax rotation
Orthogonal rotation
Oblique rotation
Factor loadings
Factor scores
Structural equation modeling (SEM)
Path diagram
Exogenous variable
Endogenous variable
Path coefficient
Direct effect
Indirect effect
Total effect
Goodness of fit
Cluster analysis
Multivariate analysis of variance (MANOVA)
Vector
Group centroid
Discriminant analysis
Canonical correlation
Wilk's lambda (λ)
Survival analysis
Terminal event
Censored observations
Kaplan-Meier estimates
Survival curve
Median survival time
Cox proportional hazards model
Hazard ratio (HR)

Chapter Summary

- Multivariate analysis refers to a set of statistical procedures that are distinguished by the ability to examine several response variables within a single analysis to account for interrelationships.
- Exploratory factor analysis (EFA) is a procedure that is used to explore the underlying structure of a set of variables related to a latent trait to reduce data to a smaller set of correlated values, or factors.
  - Factors are sets of variables that have high correlations with each other and that are not correlated with variables in other factors.
  - Factors are derived through a process of extraction that defines the number of variables that account for the majority of variance in the data, based on a variance measure called an eigenvalue.
  - A factor matrix is developed that lists factor loadings, which are the correlations of each variable with each factor. Loadings above .40 are considered meaningful to assign a variable to a particular factor.
  - A factor matrix may need to be rotated to assure a clean orientation of variables with individual factors. A rotated factor matrix will usually present more interpretable data.
  - Factors are "named" by the researcher, based on which variables are assigned to that factor.
- Principal components analysis (PCA) is a different exploratory technique that identifies *components* rather than factors that account for the total variance in a dataset.
- Confirmatory factor analysis (CFA) is used to confirm hypotheses about underlying constructs by proposing a model and then verifying it through a factor analysis.
- Structural equation modeling (SEM) is another confirmatory technique that analyzes a proposed model in terms of causal relationships that fit a given theory.
  - Data are presented in a path diagram that shows the causal and outcome relationships among latent traits.
  - Exogenous variables are those that have causes outside the model and are not included in the theory being tested. Endogenous variables are those that are caused by the exogenous variables.
  - Direct effects are those that show a direct causal connection. Indirect effects are identified by a chain of two or more variables that include a *mediating variable*, which essentially serves as both a cause and effect.
- Cluster analysis is an exploratory approach that looks at common characteristics among subgroups of individuals.
- Multivariate analysis of variance (MANOVA) is a form of ANOVA that incorporates more than one dependent variable. Its purpose is to account for shared variance that is present when dependent variables are related to each other.
  - Rather than looking at group means, the MANOVA looks at a set of means for each group, called a vector. The overall mean of the vectors is the group centroid.
  - The result of the MANOVA will indicate if groups are different based on these combined group scores.
  - Follow-up tests for the MANOVA may include individual analyses of variance, or a discriminant analysis, which looks at linear combinations of variables to determine if they can predict group membership.
- Survival analysis looks at the time to reach a *terminal event* for subgroups to determine if the risk of reaching the terminal event is different. Often used to assess risk of mortality, it can also be used to look at other outcomes.
  - Because not all participants will reach the terminal event within the time frame of the study, censored observations represent those

who have not yet reached the terminal event by the time the study ends.

- Kaplan-Meier estimates are used to determine the probability that someone will reach a terminal event within a given time period.
- The Cox proportional hazards method is used to derive a hazard ratio (HR), which is interpreted like an odds ratio. A value of 1.0 indicates no increased risk of reaching the terminal event, the null value. Values below 1.0 indicate decreased risk and above 1.0 indicate increased risk. Confidence intervals can be constructed.
- The median survival time is the point at which 50% of the participants are expected to survive, or that point at which there is a 50% chance of surviving.

CHAPTER 32

Measurement Revisited: Reliability and Validity Statistics

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss how measurement and reliability theory are related to the concept of variance.
2. Explain the purposes of the six models of the intraclass correlation coefficient (ICC).
3. Interpret computer output for ICC values.
4. Discuss how percent agreement, kappa, and weighted kappa are used to assess agreement.
5. Discuss the purpose of Cronbach's alpha in testing for internal consistency.
6. Discuss the purpose of the standard error of measurement (SEM) for understanding the reliability of clinical scores.
7. Discuss the use of Bland-Altman plots and other methods of reliability testing for parallel forms and response stability.
8. Discuss the interpretation of the minimal detectable change (MDC) and minimal clinically important difference (MCID) in the measurement of change.
9. Describe several methods for measuring change using effect size indices.

■ Key Terms

Intraclass correlation coefficient (ICC)
Random effect
Fixed effect
Standard error of measurement (SEM)
Percent agreement
Crosstabulation
Kappa statistic, κ
Weighted kappa, κ_W
Internal consistency
Chronbach's alpha
Item-to-total correlation
Limits of agreement (LOA)
Bland-Altman plot
Responsiveness
Minimal detectable change (MDC)
Minimal clinically important difference (MCID)

Distribution-based measures
Effect size
Standardized response mean (SRM)
Guyatt's responsiveness index (GRI)
Anchor-based measures
Global rating of change
MDC proportion
MCID proportion

Chapter Summary

- Reliability and validity are fundamental concepts for evidence-based practice, establishing consistency and meaning of measurements. Several statistical procedures are used to assess reliability and validity based on classical measurement theory.
  - The intraclass correlation coefficient (ICC) is primarily used to assess test-retest or rater reliability. It is a measure of relative reliability. As a unitless index, this measure allows comparison of testing methods.
  - ICC values can range from 0.00 to 1.00, with higher values indicating greater reliability.
  - The ICC has the advantage of being able to evaluate reliability across two or more sets of scores representing multiple assessments of the same test or multiple raters.
  - The ICC is calculated using variance estimates from a repeated measures analysis of variance (ANOVA).
  - There are two *forms* of the ICC, depending on whether the assigned scores are derived from single ratings or from mean ratings across several trials. Form 1 involves single ratings and form *k* a mean of several ratings. The value of *k* is the number of scores used to derive the mean.
  - There are three *models* of the ICC, depending on whether the results of the analysis are intended to be generalized beyond particular testing situation of the study, or if the reliability of those involved in the study are the only raters of interest. Only two models are typically used.
    - Model 2 is used most often, where each subject is assessed by the same set of raters who are considered "random" representatives of others who would also use the measurement. Model 3 is used when the raters are the only ones of interest, and their results are not intended to be generalized to others. ICCs are classified by model and form, such as ICC (2,1) indicating that scores are single values and will be generalized beyond the study. ICC (3,*k*) indicates that scores are means of *k* scores and will not be generalized beyond the study.
    - Researchers should decide which model fits their study's purpose at the outset. The different forms and models can generate different values, with model 2 generally being larger than model 3, and form *k* being larger than form 1.
- There are no standards for strong reliability, although measures used for clinical decision-making should be close to .90 or higher.
- The standard error of measurement (SEM) is a measure of absolute reliability, providing information about expected consistency in a measurement over several trials. It is expressed in the original unit of measure-ment.
  - The SEM represents the range of scores that can be expected when differences are

due to random error. It is calculated using relative reliability coefficients or variance measures from an ANOVA.

- The SEM is used to determine a confidence interval. For instance, the 95% CI would indicate that we were 95% confident that an individual's true score falls within a certain range, with variability due to random error. Therefore, if an individual scores 60 pounds on a test of grip strength, and we know the SEM is 4.38, we determine the 95% CI as $X \pm z$ (SEM), where $z = 1.96$. Therefore, we are 95% confident that the individual's true score would fall somewhere between 51.42 and 68.58.
- The SEM allows a clinician to determine how variable we can expect an individual's performance to be.

- Percent agreement is a measure of how frequently raters agree on categorical scores. It equals the number of observed agreements out of the total number of possible agreements.
  - Because chance can affect agreement, we use a chance-corrected statistic called kappa, κ. It represents a ratio of the number of observed agreements relative to the number of agreements that would be expected if only chance were operating.
  - Kappa will generally be smaller than percent agreement. Excellent agreement is considered above .80.
  - An alternative method called weighted kappa, κ_w, can be used when there is a difference in the importance of disagreements in different directions.
- Internal consistency is a measure of reliability that reflects the degree of consistency within a set of items on a scale, indicating the correlation among items.
  - Cronbach's alpha is the test statistic. A high value of alpha, above .70, indicates that the items on the scale relate to a single dimension. Low values of alpha may indicate that items are not measuring dimensions of the same construct.
- When we want to compare the results of different forms of a measurement we can use a procedure that looks at limits of agreement (LoA). For example, we could compare blood pressure measurements taken with manual versus automated instruments. If these are reliable alternatives, we expect their measurements to be close.
  - A Bland-Altman plot can be used to show the relationship between the two measurements, plotting the mean of the two measurements for each individual against the difference between the two measurements. A band at 1.96 standard deviations above and below the mean scores indicates the limits of agreement within which we would consider the instruments interchangeable.
- Responsiveness is the ability of a measuring instrument to register change in a score whenever an actual change has occurred. It is a ratio of true change and random error.
  - The minimal detectable change (MDC), a measure of reliability, is the minimal amount of measured change required before we can eliminate the possibility that measurement error is solely responsible. If a measurement exceeds the MDC, we can be confident that some portion of the change is due to real change in the measured variable.
  - The MDC = z * SEM *$\sqrt{2}$. The value of z indicates the levels of confidence attached to the MDC, usually 95% (z = 1.96) or 90% (z = 1.645).
  - Beyond simply looking at random error, the minimal clinically important difference (MCID) is the smallest change that is perceived as beneficial by the patient and that would lead to a change in management. The MCID reflects

validity, and is a judgment about the importance of different degrees of change.

- The MCID is usually determined by using a global measure of importance, often on an ordinal scale, such as 0 to ±5 from, with 0 indicating no change, 5 indicating a great deal better, and –5 indicating a great deal worse. Using a cutoff such as "somewhat better" or "somewhat worse," a value can be determined that matches those levels, which would be considered the MCID.
- Distribution measures can be used to assess MCID based on effect size indices. The three most commonly used are (1) the effect size index (ES), which is a ratio of the mean change score in a group of stable subjects divided by the standard deviation of baseline scores. A measure with high variability will have a smaller effect size. The standardized response mean (SRM) is a ratio of the change from pretest to posttest divided by the standard deviation of the change scores within a distribution. A distribution with high variability in the degree of change will have a small SRM. Guyatt's responsiveness index (GRI) uses an anchor-based MCID for the measure, indicating the smallest different between baseline and posttest that would represent meaningful change.

CHAPTER 33

Diagnostic Accuracy

Chapter Overview

Objectives

After completing this chapter, the learner will be able to:

1. Apply measures of sensitivity, specificity, predictive values, and likelihood ratios to the application of clinical tests.
2. Discuss the relationship among measures of diagnostic accuracy.
3. Determine pretest and posttest probabilities for a given case.
4. Discuss the application of receiver operating characteristic (ROC) curves in the interpretation of test scores.
5. Describe the use of clinical prediction rules.

Key Terms

Index test
Reference standard
Sensitivity
Specificity
True positive
True negative
False positive
False negative
False-negative rate
False-positive rate
Predictive value
Prevalence
SpPin
SnNout
Pretest probability
Posttest probability
Bayes' theorem
Likelihood ratio (LR)
Nomogram
Receiver operating characteristic (ROC) curve
Cutoff score
Area under the curve
Youden index
Clinical prediction rule

Chapter Summary

- The purpose of a study of diagnostic accuracy is to determine if a new test, the index test, is accurate based on results of a gold standard that is known to be an accurate indication of the patient's true status, either the presence or absence of the condition.
- The validity of a diagnostic test is dependent on the accuracy of the gold standard, or a

reference standard that is assumed to be an accurate criterion.

- Diagnostic tests must be evaluated by using samples that represent the full spectrum of the condition.
- Results of diagnostic study are obtained using a 2 × 2 table, with columns indicating the true diagnostic result based on the gold standard (present/absent), and rows indicating those who test positive or negative. Cells in the table use the standard *a*, *b*, *c*, *d* designations. Diagnostic accuracy is the proportion of all correct tests in the sample.
- True positives are those who are accurately diagnosed with the condition (cell *a*), and true negatives are those correctly identified as not having the condition (cell *d*). False positives are those who are incorrectly assigned a positive test result (cell *b*). False negatives are individuals who are incorrectly assigned a negative test (cell *c*).
- **Sensitivity (Sn)** is the test's ability to obtain a positive test when the target condition is really present, or the *true positive rate* ($a/(a + c)$). Specificity (Sp) is the test's ability to obtain a negative test when the condition is absent, or the *true negative rate* ($d/(c + d)$). These proportions are expressed as percentages.
- The complement of sensitivity is the false-negative rate (1–sensitivity). The complement of specificity is the false positive-rate (1–specificity). These represent the probability of incorrect tests.
- A positive predictive value (PV+) is the proportion of all those who tested positive who were true positives. A negative predictive value (PV-) is the proportion of all those who tested negative who were true negatives. Predictive value provides information regarding how useful the test will be.
- Prevalence is the number of cases of a condition existing in a given population at a point in time. This value can influence sensitivity, specificity, or predictive value if the sample does not reflect prevalence in the overall population.
- When a test has high sensitivity, a negative test rules **out** the diagnosis. When a test has high specificity, a positive test rules **in** the diagnosis. The mnemonics SpPIN and SnNout are used to remember these relationships.
- Diagnostic tests provide information about a patient's true condition, which will lead to treatment decisions. When we evaluate a patient initially we may hypothesize about a likely diagnosis. This hypothesis can be translated into a probability or level of confidence, termed the pretest probability, or *prior probability*—what we think the diagnosis will be before we perform the test. This can be estimated from clinical information or literature, or it can be estimated as the prevalence of the condition in the population or the study sample.
- Once the test is done, we can consider how much more confident we are about the diagnosis, the posttest probability, or *posterior probability*.
- When a pretest probability is high, it may not be necessary to perform a test. When it is low, it may warrant considering a different diagnosis. When there is uncertainty, the test should be done. Once it is performed, we can determine if the posttest probability is low (consider a different diagnosis), high (diagnosis is confirmed), or if there is still uncertainty, and other tests should be done.
- Likelihood ratios reflect the "confirming power" of a test, indicating how much the test can increase certainty about a positive diagnosis. A positive likelihood ratio (LR+) tells us how many times more likely a positive test will be seen in those with the

disorder than in those without the disorder [Sn/(1–Sp)]. A negative likelihood ratio (LR–) tells us how many times more likely a negative test will be seen in those with the disorder than in those without the disorder [(1–Sn/Sp]. A discriminating test will have a high LR+ and a low LR–.

  - A LR+ over 5 and a LR– lower than 0.2 represent relatively important effects. Likelihood ratios between 0.2 and 0.5 and between 2 to 5 may be important. Values close to 1.0 represent unimportant effects.

- A nomogram can be used to determine posttest probability by knowing the pretest probability and the LR+ and LR–.
- When the outcome of a diagnostic test is continuous, rather than dichotomous, a cutoff score must be set to determine who does and does not have the condition. This cutoff should provide the best balance between sensitivity and specificity.
  - This balance can be represented graphically by a receiver operating characteristic (ROC) curve, which examines the ratio of "signal" to "noise" in the data. The curve is based on the balance between Sn and 1–Sp (between true positives and false positives) at varied cutoff points. The area under the curve (AUC) can be used to determine the relative accuracy of the test. An AUC of .50 indicates a result no better than chance. The closer to 1.0, the stronger the test. The best cutoff point will typically be at the point where the curve turns.
- Clinical prediction rules (CPR) are shortcut tools to identify the likely diagnosis, prognosis, or response to treatment. A set of findings will increase the probability that a condition is present.

CHAPTER 34

Epidemiology: Measuring Risk

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the purpose of description epidemiology in relation to person, place and time.
2. Explain the difference between incidence and prevalence.
3. Calculate and interpret relative risk and odds ratios.
4. Discuss the concepts of confounding and effect modification in the interpretation of risk.
5. Explain the concepts of relative and absolute risk reduction.
6. Apply number needed to treat (NNT) and number needed to harm (NNH) to clinical data to interpret the relative effectiveness or harm of interventions.

■ Key Terms

Epidemiology
Descriptive epidemiology
Prevalence
Incidence
Cumulative incidence
Incidence rate (IR)
Person-time
Vital statistics
Mortality rate
Crude mortality rate
Cause-specific rate
Case-fatality rate
Age-specific rate
Relative risk (RR)
Odds ratio (OR)
Effect modification
Confounding
Simpson's Paradox
Analytic epidemiology
Experimental event rate (EER)
Control event rate (CER)
Relative risk reduction (RRR)
Absolute risk reduction (ARR)
Number needed to treat (NNT)
Absolute risk increase (ARI)
Number needed to harm (NNH)

■ Chapter Summary

- Epidemiology is that branch of statistics and research that focuses on the distribution and determinants of health outcomes in different populations. It can involve description to identify who, when, and where disorders occur.
- Descriptive studies will measure the incidence of a disorder, which quantifies the number of new cases during a specified time

period. It can also assess prevalence, which quantifies the number of cases existing in the population within a specified time period.

- Vital statistics indicate the rate of a condition within a population, such as birth rate or mortality rate. A crude rate is based on the total population. Rates can be adjusted to age, causes, or other categories that can differentiate outcomes.
- Analytic epidemiology is concerned with testing hypotheses, typically to determine the relationship between specific exposures and disease. Exposures may be lifestyle behaviors, occupational hazards, or environmental influences that present excess or decreased risk of acquiring a disorder.
- To assess risk, data are typically displayed in a 2 × 2 contingency table, with the columns representing the presence or absence of the disorder and rows representing the presence or absence of the exposure.
    - Cells are designated *a, b, c, d* according to typical contingency table format.
    - The incidence of a health outcome represents its absolute risk (AR). For those exposed (E), the AR_E is the number of cases among the total exposed sample ($a/(a+b)$. For those who are not exposed, AR_0 is the number of cases among the total unexposed sample ($c/(c+d)$. These estimates reflect the probability of the health outcome for the exposed and unexposed groups.
- Relative risk (RR) is the ratio of these two probabilities, indicating the likelihood that someone who has been exposed to the risk factor will develop the disorder, as compared to those who are unexposed.
    - $RR = AR_E/AR_0 = [a/(a+b)]/[c/(c+d)]$.
    - Relative risk can be used with cohort studies and with experimental studies where the exposure is the treatment condition.
    - If the absolute risk is the same for those exposed and unexposed, RR = 1.0. A RR above 1.0 indicates an increased risk, and a value less than 1.0 indicates that the exposure reduces the risk of developing the disorder (is preventive).
    - Confidence intervals can also be calculated for RR to provide a better estimate of the risk within the population.
    - In case-control studies, we cannot compute RR because the number of cases and exposed individuals are chosen by the researcher. Therefore, we use an estimate of RR called the odds ratio (OR), which is equal to ad/bc. The odds ratio is interpreted in the same way as RR.
- Effect modification refers to a difference in the magnitude of an effect measure across levels of another variable. An effect modifier will change the relative risk associated with an exposure for different subgroups in the population, a form of interaction.
- A confounding variable is a nuisance variable, a third variable that is associated with the exposure, is a risk factor for the disorder independent of the exposure, and is not part of the causal link between exposure and disease.
- Analytic epidemiology can also be applied to measures of intervention effect in the format of an RCT, with an experimental and control group. By setting a threshold for a successful outcome of treatment, it is possible to determine how much more likely this success is seen above "failed" outcomes.
    - The experimental event rate (EER) is the proportion of people in the treatment group who experience an adverse outcome (an adverse event is considered a failure of treatment). The control event rate (CER) is the proportion of people in the control group who experienced an adverse outcome.

- The ratio of these two values is the RR associated with the intervention = EER/CER. These values are essentially the same as the measure of absolute risk.
- The relative risk reduction (RRR) indicates how much the risk is reduced in the treatment group as compared to the control group = (CER–EER)/CER.
- Absolute risk reduction (ARR) indicates the actual difference in risk = CER – EER.

- The degree of risk reduction should help us decide whether a treatment is worth pursuing based on the likelihood of a successful outcome, or avoidance of an adverse event if the purpose of treatment is prevention. As a useful clinical value, the number needed to treat (NNT) is the number of patients that would need to be treated to prevent one adverse outcome or to achieve one beneficial outcome in a given time period.
  - NNT is the reciprocal of the ARR (1/ARR).
  - An NNT of 1.0 means that we would need to treat 1 patient to avoid one adverse outcome; that is, every patient will benefit from treatment. An NNT of 10 means that we would need to treat 10 patients to prevent one adverse outcome, or 1 out of 10 will be successful. The lower the value of NNT, the more effective the treatment.
- Where NNT reflects the prevention of an adverse outcome, we must also be concerned about potential harms.
  - The absolute risk increase (ARI) is the amount of increased risk associated with an intervention, or EER – CER.
  - The number needed to harm (NNH) is the reciprocal of ARI (1/ARI). It is the number of patients that would need to be treated to *cause* one adverse outcome.
  - The larger the NNH, the less likely a patient is to experience an adverse outcome. An NNH of 100 would mean that we would need to treat 100 patients before we would see 1 adverse outcome. An NNH of 1.0 would mean every patient would have an adverse outcome.
- Several factors must be considered in comparing values of NNT across studies:
  - The NNT must be interpreted in terms of the time period for treatment and follow-up.
  - The interpretation of NNT will depend on baseline risk.
  - To compare NNTs across studies, outcomes must be the same.
  - The design validity of the trial must be considered.
  - There are no standard limits to indicate how NNT and NNH should be used for decision-making. Decisions need to consider severity of the disease and its consequences, availability of treatments, side effects, cost, and patient preferences.

CHAPTER 35

Writing a Research Proposal

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Explain the purposes of a research proposal.
2. Describe the components of a research proposal.
3. Describe the important issues for administrative support of a research proposal.
4. Describe how research personnel may be involved in the development of a research proposal.
5. Develop a timeline for a research study.

■ Key Terms

Abstract
Specific aims
Objectives
Hypotheses
Principal investigator
Timeline
Budget
Direct costs
Indirect costs

■ Chapter Summary

- The initial stages of the research process include development of the research question and delineation of methods, data collection, and data analysis procedures. To be sure that these plans are carried out, a research proposal describes the purpose of the study, the importance of the research question, the research protocol, and feasibility of the project.
  - The research proposal contains a synthesis of the scientific literature that supports the research question and provides a theoretical rationale for the study and its methods.
  - Proposals may serve as the body of a grant application.
  - Proposals will be reviewed by IRBs to assure that all procedures are ethical, that the project personnel are qualified, and the project is feasible.
  - Proposals serve as a blueprint for all involved in a study to be sure they understand their roles and how procedures should be followed to assure consistency and accuracy throughout the project.

- The first part of a proposal is the research plan.
  - The proposal begins with its title, which should reflect the study fully, and an abstract that summarizes the intent of the study. These are often written after the study is completed.
  - The opening section of a proposal identifies the research question, why it is important, how it is supported by literature or theory, and what contributions the study will make.
  - The purpose of a study is usually stated in terms of hypotheses, guiding questions, or specific aims that reflect the expected outcomes of the study.
  - The methods section should include a description of the study design, which may be experimental, observational, descriptive, or qualitative.
  - It will specify who participants will be (inclusion and exclusion criteria), how and from where they will be selected, how informed consent will be obtained, and how sample size was determined. The informed consent form will be included in the proposal for IRB review.
  - Procedures are fully explained as they will be carried out for data collection, including details on devices used and full operational definitions.
  - A timeline flow sheet is useful to summarize the expected sequence of events for the planning and implementation of a study.
  - Data management is described in terms of analysis procedures, safety and confidentiality of data.
  - Appendices may follow the main section, including sample surveys, equipment details, letters of support from outside agencies, or other information that is necessary to support the project's implementation.
- The second part of the proposal is the plan for administrative support.
  - This includes a full description of the investigators, their qualifications, and other personnel involved in the study.
  - Information must be provided that demonstrates how existing resources are available to support the project or what additional resources will be necessary and how they will be obtained.
  - A full budget for direct costs must be detailed for funding agencies but may be required for academic proposals as well. This should include personnel salaries and the proportion of time the individuals will devote to the study, equipment costs, facilities, and supplies. It will also include indirect costs that relate to overhead, usually at a fixed rate for funding agencies.
- The format of a proposal is stipulated by the agency to which it will be submitted and should be followed exactly.
- Proposals should be written clearly, with concise language, and using an organizational structure that will make it easy for reviewers to follow.
- Agencies will have specific requirements for submitting proposals, including deadlines that must be followed.

CHAPTER 36

Critical Appraisal: Evaluating Research Reports

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Discuss the role of levels of evidence in critiquing research.
2. Describe the core questions that can be used to assess all types of research studies.
3. Discuss the components of a research report and the basic information that should be included to determine if a study is valid.
4. Discuss the information that is necessary to determine if results of a research study are meaningful.
5. Describe specific questions that are relevant to critically appraise studies of interventions, diagnosis, prognosis, or qualitative research.
6. Discuss the purpose and format of a *critically appraised topic* (CAT), and create a CAT for a given research report.

■ Key Terms

Evidence-based practice
Critical appraisal
Levels of evidence
Critically appraised topic (CAT)
Journal club

■ Chapter Summary

- The purpose of critical appraisal is to determine the scientific merit of a research report and its applicability to evidence-based practice.
- The classification of levels of evidence can be used as an initial criterion to judge the rigor of a study's design. Use of a strong design, however, does not guarantee the validity of a study's findings.
- The appraisal process begins with a search to identify a relevant study to answer a clinical question. The title and abstract will be the first pass at determining if the study is appropriate.

- Studies must be evaluated based on the background presented to justify the research question, the validity of the methods used, how the results are presented, and how findings are put in context of prior research and clinical theory.
- Several core questions can be used for any type of study to answer the questions, "Is the study valid?" "Are the results meaningful?" and "Are the results relevant to my patient?"
- Studies must present information to establish the study's purpose and rationale, who the participants are and how they were selected, what study design was used, how data were collected and analyzed, what results were obtained, and how findings can be interpreted.
- Studies should present findings in terms of effect size so that the importance of the findings can be determined.
- Studies of intervention should include information about how treatment groups were formed, operational definitions of intervention and measurement outcomes, procedures to eliminate bias such as blinding and random assignment, and use of appropriate analysis procedures, including intention to treat analysis.
- Studies of diagnostic accuracy should include information on the gold standard used to assess the index test, procedures to eliminate bias such as blinding, and how measures of diagnostic accuracy were applied.
- Prognostic studies should include information regarding the retrospective or prospective nature of the design, how cases and controls were defined and recruited, how data were obtained, the relevant time period, and how prognostic estimates were derived.
- Qualitative studies should include information on the sampling strategy and the qualitative approach used, such as ethnography, phenomenology, or grounded theory. Data collection and analysis procedures should be described in detail, including how trustworthiness was assured.
- A critically appraised topic (CAT) is a brief summary of one or more research reports that focus on a specific patient question. Its sections include a title that reflects the question, the author and date, a clinical scenario describing the patient case, the clinical question using the PICO format, a clinical bottom line that summarizes the applicability of findings, the search history and citations, a summary of the study's design and the results, and other comments that can address internal and external validity.
- Journal clubs can provide a useful venue for teaching practitioners to search for articles and to develop critical appraisal skills.

CHAPTER 37

Synthesizing Literature: Systematic Reviews and Meta-Analyses

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the purposes and formats of systematic reviews, meta-analyses, and scoping reviews.
2. Discuss how publication bias and grey literature can influence outcomes of a systematic review or meta-analysis.
3. Discuss how selection criteria are applied in the development of a systematic review.
4. Describe the major components in a search strategy to develop a systematic review.
5. Apply scales for assessing methodologic quality of research studies.
6. Discuss how effect size is used to compare studies for meta-analysis.
7. Interpret data from a forest plot and funnel plot.
8. Describe the purpose of sensitivity analysis.
9. Discuss the criteria for appraising systematic reviews and meta-analyses.
10. Describe the purpose of scoping reviews and how they are distinguished from systematic reviews.

■ Key Terms

Systematic review
Meta-analysis
Scoping review
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)
Publication bias
Grey literature
Data extraction
Template for Intervention Description and Replication (TIDieR)
Methodological quality
Enhancing the QUAlity and Transparency OF health Research (EQUATOR)
Physiotherapy Evidence Database (PEDro) scale
Cochrane risk of bias tool
Selection bias
Performance bias
Attrition bias
Detection bias
Quality Assessment of Diagnostic Accuracy Studies (QUADAS)
Quality in Prognosis Studies (QUIPS)
COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)

Cochrane Handbook
Effect size
Standardized mean difference (SMD)
Forest plot
Heterogeneity
Cochran's Q
I^2 statistic
Sensitivity analysis
Funnel plot
Grade of Recommendation, Assessment, Development and Evaluation (GRADE)
A MeaSurement Tool to Assess systematic Reviews (AMSTAR)
PRISMA Extension for Scoping Reviews (PRISMA-ScR)

Chapter Summary

- Systematic reviews are a form of research that uses a rigorous process of searching, appraising and summarizing existing literature on a selected topic. When effect size data are available in research reports, a meta-analysis can be done by pooling effect size estimates, creating a more robust estimate of an intervention effect.
- Systematic reviews commonly focus on interventions but can also be applied to studies of prognosis, diagnosis, and qualitative studies.
- Guidelines for reporting systematic reviews are included in the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) checklist.
- A systematic review is built on a specific research question that should include the elements of PICO: population, intervention and comparison, and outcome measures.
- The process of developing a systematic review is similar to the typical research process, starting with an introduction with background literature and a research rationale.
- The methods section details the criteria for choosing studies to review, describes the process of searching for references, and how the studies were selected and evaluated. In a systematic review, the "subjects" of the study are the articles that are included. Therefore, inclusion and exclusion criteria indicate the design, participants, and definitions of interventions and outcomes that must be present in selected studies.
- The search strategy is important to assure that selected studies are comprehensive and represent the most recent research on the topic. Reviewers may restrict articles to randomized trials but should also explore studies with other designs that may still provide useful information.
- Grey literature is composed of nonpublished information, often on websites, in reports, conference proceedings, theses or dissertations. These sources can be valuable to a full understanding of the scope of knowledge on the topic.
- Reviewers want to avoid the impact of publication bias, which is the tendency for studies with "no significant" findings not to be published.
- Once articles are chosen, they must be reviewed to determine if they fit the set criteria. Data extraction is a process whereby information is analyzed and recorded as part of the review.
- The PRISMA flow diagram demonstrates how articles were accessed, and which were rejected for review. All systematic reviews should be completed by at least two colleagues who review studies independently. Reliability of their evaluations should be assessed.
- An important aspect of systematic reviews is the assessment of methodological quality, to determine if there is adequate control or sources of bias that affect the value of the results for answering the research question.

Several scales have been developed for quality of assessment of specific types of designs.

- For RCTs the *Physiotherapy Evidence Database (PEDro) scale* and the *Cochrane Risk of Bias Tool* are often used.
- For observational studies, the *Newcastle-Ottawa Scale (NOS) Quality Assessment* is commonly used.
- For diagnostic studies, the *Quality Assessment of Diagnostic Accuracy Studies* (QUADAS) is used.
- The *Quality in Prognosis Studies* (QUIPS) scale is used for prognosis studies.
- The *Consensus-based Standards for the Selection of Health Measurement Instruments* (COSMIN) is a checklist for use in methodological studies.

- Discussion and conclusions in a systematic review describe the completeness of the evidence, overall quality, and consistency of findings with other studies or reviews
- Meta-analysis is an extension of a systematic review that incorporates statistical pooling of data across several studies.
  - Pooling is based on effect sizes, such as the standardized mean difference for interventions or relative risk for observational studies.
  - Effect sizes are weighted based on sample size.
- A forest plot is a graphic display of weighted effect sizes of individual studies with confidence intervals, and a cumulative summary effect size.
- Measures of heterogeneity are applied to data to reflect consistency of findings across studies.
  - Cochran's Q and the I^2 statistics are used to represent the percentage of variance across studies that is due to true differences in treatment effect.
- Sensitivity analysis involves looking at subsets of data to see if results agree with the overall data.
- Funnel plots are scatterplots of treatment effect sizes against a measure of variability within the data for each study. If studies are not equally distributed around the total effect, it may indicate publication bias.
- The *Grade of Recommendation, Assessment, Development and Evaluation* (GRADE) system can be used to assess the quality of the body of evidence, rating quality from high to very low.
- *A Measurement Tool to Assess Systematic Reviews* (AMSTAR) is a checklist for appraisal of systematic reviews.
- Scoping reviews are a different form of systematic review that does not include the same rigor of assessment. Its intent is to address specific clinical questions to explore the nature of evidence.
  - The PRISMA *Extension for Scoping Reviews* (PRISMA-ScR) includes a checklist that defines standard reporting items for scoping reviews.

CHAPTER 38

Disseminating Research

Chapter Overview

■ Objectives

After completing this chapter, the learner will be able to:

1. Describe the components of a research article and their associated content.
2. Discuss the relevant information that should be included in each section of a research report.
3. Discuss the use of tables and graphs in a research report.
4. Assess examples of good writing style.
5. Compare and select options for dissemination.
6. Describe the elements of good design for a research poster.
7. Describe the important considerations in developing an oral presentation.

■ Key Terms

EQUATOR (*Enhancing the QUAlity and Transparency Of health Research*)
CONSORT (*CONsolidated Standards of Reporting Trials*)
Journal impact factor
Open access journals
Think. Check. Submit.
Peer review
Primary author
Corresponding author
Fabrication
Falsification
Plagiarism
Systematic reviews
Meta-analysis
Perspectives/Commentaries
Short reports
Letters to the editor
Trial protocols
Book reviews
Editorials
Oral presentations
Poster presentation
e-poster

Chapter Summary

- For research results to inform evidence-based practice, they must be disseminated in the scholarly literature. The most traditional vehicle for dissemination is a journal article.
- To select which journal to target your manuscript, consider its aims and scope, impact factor, and alternative metrics of citation.
  - The impact factor measures how frequently a journal's articles have been cited elsewhere in the previous 2 years.
- Open access journals charge authors for publication and make articles available to readers for free.
  - *Green open access* means an article is published in a traditional journal and is made available in a separate repository.
  - *Gold open access* refers to journals that make all articles freely available.
  - *Hybrid gold access* refers to traditional subscription journals that provide an option for authors to publish with free access.
  - Some journals, considered predatory journals, take advantage of authors by charging for publication but offer no rigorous peer review or editorial services.
- The International Committee of Medical Journal Editors (ICMJE) has established four criteria for authorship. All authors on a manuscript should meet all four:
  - Substantial contribution to conception and design of the work or the acquisition, analysis or interpretation of data.
  - Drafting the work or revising it critically for important intellectual content.
  - Final approval of published version .
  - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
- Most journals require all authors to sign a copyright release that includes a statement of their contributions to the work, including:
  - Conception and design
  - Participation in data collection or analysis
  - Writing the article
  - Project management or obtaining funding
  - Reviewing the final paper
- The *primary author* takes responsibility for coordination of the team involved in the publication. This person is usually also the *corresponding author*, who takes responsibility for communication during submission and revision.
- Guidelines for reporting research results provide a checklist for what to include in a manuscript. Guidelines have been developed for many kinds of research designs, and they can be accessed via the EQUATOR network. Most guidelines also have *Elaboration and Explanation Papers* that clarify each item on the checklist.
- Journal articles follow the IMRAD format: introduction, methods, results, and discussion. Many journals also allow online supplementary material.
- Using clear language with active verbs and concrete nouns helps readers grasp the importance of your research. Avoid abbreviations and jargon that may cloud meaning.
- Other vehicles for presenting research results include case reports, short reports, letters to the editor, and oral and poster presentations.

- With the profusion of publications, authors need to promote their work to have it reach their intended audience. Social media can help spread the word about articles and provide additional measures of their impact.