THE RESEARCH PROCESS: DESIGN
CHAPTER OVERVIEW

This chapter covers research design and methodology that become the methods section of a research project. Choosing the right design adds validity, reliability, and strength to the results and conclusions. The best of projects can be derailed by errors in experimental design and weakened by a poor choice of research design. Challenges to validity and reliability that may emerge at the peer-review stage must be considered when developing the methods for an investigation.

Methodology

Christopher E. Bork, PhD
Robert W. Jarski, PhD, PA-C
J. Glenn Forister, MS, PA-C

Learning Objectives

• Define research methodology.
• List the elements of methodology.
• Draw the schematic for selected research designs.
• List, define, and discuss the type of analysis for each research design.
• List and discuss the types of studies.
• Define, identify, and discuss the terms bias, error, reliability, validity, threats to validity, and types of data.

INTRODUCTION

Once the topic and specific research questions, or hypotheses, have been identified and the literature review has been completed, the investigator must choose a methodology suitable for achieving the project’s objective. This objective is stated as a hypothesis, a purpose statement, or a clear and specific research question. Additionally, the primary objective should be evaluated by this criterion: Is it important to medicine, the profession, or society? That is, will the study make a significant contribution to the field’s body of knowledge? Some journals refer to this as the “so what?” question. In many cases, authors are required to respond in writing to this question when submitting a manuscript for publication. Answering this question helps the investigator condense the meaning and value of the
The research methods should be driven by the research objectives. When choosing a methodology, two mistakes may distract novice investigators. The first mistake often made is selecting a familiar methodology without first defining the research objectives. The second most common mistake is planning to use already available data. Although convenient, this backward approach may produce trivial information that is unlikely to contribute to the field. The methodology should generate data that achieve worthwhile objectives.

### Table 6-1 Selected Research Designs

<table>
<thead>
<tr>
<th>Name</th>
<th>Design</th>
<th>Statistical Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-shot case study</td>
<td>X O</td>
<td>None</td>
</tr>
<tr>
<td>One group pretest–posttest</td>
<td>O X O</td>
<td>Dependent or paired t-test</td>
</tr>
<tr>
<td>Static group comparison</td>
<td>X O</td>
<td>Independent t-test</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>Chi-squared</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mann-Whitney U</td>
</tr>
<tr>
<td>Posttest-only control group</td>
<td>R X O</td>
<td>Independent t-test or ANOVA</td>
</tr>
<tr>
<td></td>
<td>R O</td>
<td>Mann-Whitney U</td>
</tr>
<tr>
<td>Nonequivalent control group</td>
<td>O X O</td>
<td>ANCOVA or ANOVA</td>
</tr>
<tr>
<td></td>
<td>O O</td>
<td></td>
</tr>
<tr>
<td>Pretest–posttest control group</td>
<td>R O X O</td>
<td>ANCOVA or ANOVA</td>
</tr>
<tr>
<td></td>
<td>R O O</td>
<td></td>
</tr>
<tr>
<td>Solomon four group</td>
<td>R O X O</td>
<td>ANCOVA or ANOVA</td>
</tr>
<tr>
<td></td>
<td>R O O</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R X O</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R O</td>
<td></td>
</tr>
<tr>
<td>Counterbalanced</td>
<td>X₁ O X₂ O X₃ O</td>
<td>ANOVA</td>
</tr>
<tr>
<td></td>
<td>X₂ O X₃ O X₁ O</td>
<td></td>
</tr>
<tr>
<td></td>
<td>X₃ O X₁ O X₂ O</td>
<td></td>
</tr>
<tr>
<td>Time series</td>
<td>O O O O O X O O O O</td>
<td>ANOVA, trend analysis</td>
</tr>
</tbody>
</table>

X = experimental treatment or intervention; O = observation, measurement, or evaluation; R = randomization of a large number of subjects; ANOVA = analysis of variance; ANCOVA = analysis of covariance.

### THE METHODS SECTION

The methods section describes how the research study was conducted. It should be sufficiently clear and detailed so that others can duplicate the study. The methods section includes descriptions of the following:

- Subjects
- Instrumentation used (including questionnaires, when applicable)
- Procedures performed
- Analytic procedures used for evaluating and summarizing the data
Data analysis must be suited to the particulars of the study: the number of subject groups (e.g., two: an intervention and a control group), the number of subjects within each group, and the procedures used for data collection (e.g., single or repeated measures). Many healthcare researchers value a quantitative analytic approach using statistical methods. Although this is often the case, some questions in health care must be answered using a qualitative approach.

**A Methods Scheme**

Each study plan poses its own challenges. A system such as the one presented in Table 6–1 is a useful tool for identifying some study options. First, the researcher must decide whether the approach will be analytic or descriptive by answering the question: Is there a comparison between groups in this study?

**UNDERSTANDING RESEARCH DESIGN**

**Controlling Bias**

When designing research, the investigator must limit or control factors and biases that could potentially contaminate a study. As mentioned in previous chapters, all research starts with an idea or a problem. The research hypothesis is the investigator’s expectation for the outcome or the solution to the problem. The objective of a research study is to prove or disprove the investigator’s hunch.

When selecting a research design, the investigator must consider the expectation (i.e., the research hypothesis) and create a means of controlling biases that may result from this expectation. In other words, an informed, objective observer must be able to conclude that the researcher’s bias (or prior expectations) did not influence the results of the study. The investigator must design the research to eliminate bias, thus allowing the results of the study to truly represent the effect of the independent variable.

Other reasons for understanding research design include recognizing and minimizing the effect of threats to validity, both internal and external. In this chapter, the concepts of error and reliability, forms of validity, and a method of identifying and diagramming typical research designs will be discussed.

**Error**

All research involves measurement, and all measurements involve error. The common formula for a given measurement is

\[
\text{Observed measurement} = \text{True measurement} + \text{Error}
\]

As the concept of “error” decreases, the observed measurement begins to approximate the true measurement. In everyday life, errors in measurement are common; for example, anyone who has ever tried to cut a shelf to put inside a closet may be painfully aware of the formula for an observed measurement. If the observed measurement includes too great of an error, then the shelf will be either too large or too small and will not fit in the closet. Therefore, the error term can be either positive (too large) or negative (too small).

**Types of Error**

Error can be categorized into two forms, random and systematic. Random error consists of those errors that occur strictly because of chance; these errors are often thought of as “noise in the system.” Small sample sizes tend to be more vulnerable to random error. For example, if five measurements are taken versus one measurement, the average of the five measurements is less likely to be incorrect or have a large error term. The carpenter’s proverb of “measure twice, cut once” affirms the role of random errors.

Systematic error can be thought of as a series of consistent biases affecting a measurement. Typical researcher errors may be related to poor technique, such as sloppiness, or to inappropriate protocols or research designs, inappropriate measures, or incorrect statistical applications. For example, if a researcher is using heart rate (beats per minute) as an indicator and calculates heart rate using 10-second readings in some cases and 15-second readings in others, then the error risk increases. The measuring technique introduces errors of inconsistency.
Similarly, if a researcher uses a survey to assess a clinical outcome, error may be introduced by several factors including differences in verbal fluency among subjects or even the conditions under which the survey is completed. For instance, consider the differences in response to a telephone survey about practitioner satisfaction by an individual who has just experienced a 2-hour traffic jam versus a responder who has just exercised and feels wonderful. Differences in their levels of stress may affect their responses.

By choosing appropriate measurements, using reliable techniques, and employing valid instruments to obtain measurements, a researcher can eliminate a substantial number of experimental errors in data collection. Reliability and validity are fundamental to obtaining appropriate and useful data.

**Reliability**

Reliability focuses on the consistency with which a measurement is taken. If a measurement lacks reliability, then the data obtained may be useless because of error. In other words, if substantial error exists in the measurements, then the researcher cannot know whether observed changes in the dependent variable are caused by manipulation of the independent variable or by poor measurement. Reliability is also of paramount importance to the professional in clinical practice. If a clinician does not gather reliable data, then there is no way of knowing whether apparent changes in the patient are the result of actual physiological changes or poor technique. In other words, the clinician will not know if progress is taking place as a result of treatment.

**Forms of Reliability**

There are three common forms of reliability: instrument reliability, intrarater reliability, and interrater reliability. Instrument reliability indicates the consistency of measurement by a particular instrument. For example, if a weighing scale has a worn spring, it may measure lighter weights accurately but heavier weights inaccurately. The method to improve instrument reliability is to consistently calibrate the instrument.

Intrarater reliability indicates the consistency with which an individual takes measurements. For example, does the healthcare provider measure blood pressure in the same way each and every time? If not, there is a strong possibility that the measurements will differ because of technique rather than differences in the patient. One method of improving intrarater reliability is to consistently follow an established protocol and to routinely check for consistency.

Interrater reliability indicates the consistency in measurements among individuals taking the measurements. If more than one individual is taking a measurement, there must be adequate assurances that any changes are caused by changes in the true measurement, rather than fluctuations in human error. The concept of interrater reliability is often ignored and has important ramifications for clinical research. Inattention to interrater reliability may obscure differences in the research outcome. If individuals collecting data do not perform the related tasks in the same way, the researchers will not be able to ascertain whether differences are caused by differences in the patient or differences in the way the measurement was performed (i.e., an error). In other words, if data are not gathered in a consistent or reliable manner, their accuracy is questionable and, therefore, the data may be useless.

The “two P rule” (protocol and practice) applies when two or more individuals are taking measurements. Measurements should be taken using a standard protocol that has been practiced. Data collectors should also compare their measurements periodically; they may even wish to determine their consistency by performing one of the tests that assess interrater reliability, such as the kappa-statistic test. When reviewing any article that involves measurements, the reader should look for an assessment or other assurance that the data were gathered reliably.

**Validity**

Another concept that must be considered regarding data and measurement in an investigation is validity. Validity indicates the usefulness or
appropriateness of the data being gathered. In a practical sense, reliability and validity are related concepts. Reliability focuses on consistency of measurement, whereas validity focuses on the appropriateness of a given measurement. A simple illustration is a game of darts. Reliability can be viewed as the consistency of the pattern or spread of the darts. Validity can be thought of as ensuring that the darts are aimed at a dartboard. There are two principal types of validity: measurement or test validity, and design or experimental validity.

Measurement or Test Validity

Measurement or test validity answers the question: Does the test or measure actually do what it is intended to do? For this to occur, a given measurement should have a defined purpose and should relate to a given phenomenon (e.g., a clinician routinely takes a temperature because this vital sign can be an indication of an inflammatory process). For the purposes of this chapter, we will focus on some common forms of test validity.

The first form is face validity, which addresses the question: Does the particular measurement or method appear to be appropriate? This form of validity often relies on the opinion of experts. Most authorities consider face validity the weakest form of test and measurement validity. Another form is construct validity, which assesses the degree to which the measurement is based on theory. In the example of measuring body temperature, the construct that an inflammation involves heat provides modest construct validity. On the other hand, content validity asks whether the test is broad enough to address the scope of the content. For example, if students’ knowledge of anatomy were being tested but they were only tested on the anatomy of the upper extremity, then that particular test would lack content validity. Finally, criterion validity is an indication of how well the test performs and whether it is useful when judged against a standard. There are generally two subcategories of criterion validity: predictive validity and concurrent validity. Predictive validity assesses whether and how well a test predicts a specific phenomenon or outcome. For example, how well does a positive straight leg raise accurately predict a lumbar disc protrusion? Concurrent validity asks whether the test performs as well as an accepted test. Generally, this category is used to validate a short or noninvasive version of a test. For example, concurrent validity would be used to establish the validity of a urine test, as opposed to a serum glucose test, to monitor diabetes mellitus.

Design or Experimental Validity

There are two forms of design or experimental validity: internal and external. Internal validity is concerned with limiting or controlling factors and events other than the independent variable, which may cause changes in the outcome, or dependent variable. These factors or events are known as threats to internal validity. External validity, on the other hand, is concerned with factors that may affect the generalizability of the conclusions drawn from the study. These factors are referred to as threats to external validity. The next section examines these two concepts.

THREATS TO INTERNAL VALIDITY

As mentioned, threats and concerns related to internal validity are unintended factors and conditions that can affect the results. For example, if an investigator is assessing the effects of two dietary regimens and does not take into account the subjects’ levels of activity, then the internal validity of the study is threatened.

There are two broad categories of threats to internal validity: temporal or time-based effects, and measurement effects. Temporal or time-based effects consist of history, maturation, or attrition. History refers to effects on the dependent variable that are a result of the passage of time. For example, suppose an investigator is interested in a new topical ointment for the common cold sore caused by the herpes simplex virus. The researcher treats
one group of patients with a new drug or topical ointment for 7 days and obtains excellent results. However, cold sores from the herpes simplex virus are thought to be self-limiting anyway, with symptoms generally resolving in 7 days. Therefore, the passage of time has obscured the effect or noneffect of the topical ointment.

The next temporal effect or threat to internal validity is maturation, which can be thought of as threats that happen by changes as a result of development. Suppose an individual suggests that a particular type of rehabilitative therapy improves the development of infants’ motor skills. In this example, the individual contends that the therapy helps infants walk sooner. The effects of that therapy and potential changes in the infants’ motor skills may not be the result of the type of therapy involved, but rather of the developmental process. Thus, the experiment may be flawed by threats to internal validity, specifically maturation.

The third temporal effect or threat to internal validity results from attrition. When subjects leave a study prematurely, the results may be distorted. Consider the consequences of a study involving a new drug for migraine headache sufferers. In this hypothetical study, one group of subjects receives the new drug, and their results are compared with another group of subjects who receive a placebo. The subjects will be seen every 3 weeks for a period of 6 months. Suppose subjects taking the new drug no longer experience migraine headaches, so they no longer come to their appointments. The only participants left in the study are those for whom the drug did not work. Therefore, when comparing the placebo to the experimental drug after 6 months, there is apparently no difference between the groups because the majority of participants who continue to have migraine headaches remained in the study. The subjects who were helped by the drug dropped out of the study, thereby affecting and possibly even distorting the results.

Measurement effects are those threats to internal validity that result from an investigator trying to measure a phenomenon. The first threat is testing, especially when the test is repeated several times. Sometimes the actual act of performing a test on a patient affects the results of the study. Consider a hypothetical investigation in which the researcher is interested in the effect of a particular type of setting on function in individuals who have suffered a cerebral vascular accident. One of the measurements may be a functional test (e.g., how well the individual is able to dress without help from others). In this example, the investigator decides to administer a pretest to determine the baseline time needed for an individual to self-dress. In this case, having the patient get dressed may help the patient discover new and better strategies for getting dressed, thus contaminating the results. As a second example, consider an investigator who wishes to compare two forms of drug therapy on patients with cardiac problems. In this hypothetical study, the investigator chooses to use a step test (i.e., have the subjects step up and down repeatedly) and record the number of times subjects can continue until the heart rate reaches a predetermined maximum percentage. In this study, the patient’s performance may be affected by the motor learning that takes place. In other words, the patient’s coordination may improve just by virtue of a pretest using the step test.

The next threat to internal validity is instrumentation. The type of instrumentation used may affect the results. Consider a study in which the investigator is interested in whether children with handwriting problems press their pencils harder on the paper. To measure the point pressure of the writing implement, the children are asked to use a pencil-type instrument containing a force transducer and a wire that leads to a recording device. The fact that the instrument represents an unnatural pencil may affect the results. Thus, the instrument itself is a threat to internal validity.

Another threat to internal validity is sampling. Sampling effects include the selection of subjects for a study according to some bias, whether recognized or not. If selected by virtue of a bias, the subjects are not representative of the population. For example, if subjects were surveyed in a study via a mail questionnaire that was sent only to residents in an affluent suburb, then the conclusions drawn
from the results may be affected by the sample that was surveyed.

The final measurement effect that represents a threat to internal validity is statistical regression to the mean. Simply stated, this is the tendency for a group of outliers to move toward the mean (the average), not necessarily because of any difference in the subjects’ characteristics, but because of the laws of probability. A classic example of statistical regression to the mean can be illustrated by what happened when students with developmental and learning challenges were mainstreamed into a classroom with so-called “normal” students. After a period of time, they were tested and their scores on a developmental and learning inventory improved. The researchers then hypothesized that, if scores improved by putting challenged students in a “normal” class, then perhaps putting “normal” students in a class with gifted students would bring the “normal” students’ aptitudes up. However, when the students in the class were retested, the exceptional students appeared to do worse on the test. Did the exposure to “normal” students somehow contaminate these learners? The answer is no. What was happening was simply statistical regression to the mean. In both cases, the group that was being tested (the developmentally challenged and the gifted students) came from the ends of the distribution of test scores, that is, the highest and lowest scores. On the retest, their scores tended to migrate toward the mean or toward that typical score within a population. Therefore, investigations may be affected by statistical regression to the mean by utilizing subjects who may be considered outliers.

THREATS TO EXTERNAL VALIDITY

As previously mentioned, threats to external validity include factors and conditions that affect the ability to generalize the results of a study. Threats to external validity can be placed into two categories: threats related to the populations used and those related to the environment in which the study takes place (i.e., environmental threats).

Population-Related Threats to External Validity

The first population-related threat to external validity concerns the subjects’ accessibility to the study. When one performs a study, one usually studies a portion of a given population—a sample. If the sample used in an experiment or investigation is substantially different from the population, then the ability to generalize the results to the population may be compromised. This threat may be of particular interest in clinical studies. In the majority of clinical studies, the subject generally has access to medical care and the ability to continue with treatment. Therefore, these subjects may not represent the entire population, which may include individuals who have compromised access to medical care.

The second threat to external validity in the population category is known as subject–treatment interaction. This threat can be described as the confounding effects of the subjects’ attributes on the dependent variable. Because of genetic makeup, lifestyle, or other confounding variables, certain subjects react differently (i.e., either more positively or more negatively) to any given treatment. For example, some people are blessed with a metabolism that allows them to eat whatever they wish without gaining any weight. On the other hand, other people who follow the same diet may gain weight.

Environmental or Experiment-Related Threats to External Validity

One environmental or experiment-related threat to external validity is the description of the variables. If the variables used in a study are not described precisely and with sufficient detail, then it may be difficult for subsequent investigators to replicate the study or obtain the same results. For example, consider a new antihypertensive drug study where the control subjects receive “conventional
therapy," but this is not specified in detail. A clinician may not be able to determine whether the new drug is preferable to current therapy if not enough detail is available to ascertain how similar the new drug therapy is to the control therapy. A second environmental threat takes place when there are multiple treatments involving test order. In some cases, when two treatments or two drugs are administered, one may potentiate the effects or affect the actions of the other. Failure to recognize the effect of treatment order or multiple treatments may hinder the practitioner’s ability to apply the results of the study.

Named after a classic experiment by Elton Mayo, the Hawthorne effect is a threat to external validity. Mayo was looking at worker productivity at the Hawthorne Works plant located outside of Chicago. Mayo was interested in the effect of lighting on worker productivity. He explained to the workers that they would be in an experiment on productivity. Mayo then proceeded to increase the ambient lighting in the factory. As anticipated, productivity improved. In the next phase of his study, Mayo then dimmed the ambient lighting. Worker productivity increased again. Finally, Mayo raised the ambient lighting to the previous high-intensity level, and yet again productivity increased. Mayo concluded that the ambient lighting was unrelated to the workers’ performance—rather, the fact that the workers knew they were being studied affected productivity. The Hawthorne effect is an effect on results caused by the subjects’ knowing that they are participating in an experiment. Typically, subjects in clinical experiments have better compliance with treatment regimens than patients who are not participating in an experiment.

The Rosenthal effect refers to results caused by the involvement of the investigator in a study. The personal attributes, charisma, and abilities of the researcher can affect the results. For example, if an investigator is a charismatic practitioner, then patients may improve partially because of their belief in the clinician treating them. Other practitioners who attempt to obtain similar results are likely to be unsuccessful.

In every research study, the investigator must consider and try to control factors, conditions, and the effects of circumstances that threaten both internal and external validity. Similarly, professionals critically reading the literature must be aware of these threats to validity in order to determine the credibility of the conclusions and the applicability of the work in their practice. In many cases, the choice of research design affects the investigation’s susceptibility to threats to validity.

**TYPES OF STUDIES**

**Descriptive Studies**

Descriptive studies generate data that are either numerical or non-numerical. Non-numerical information may be presented as verbal commentaries about subjects’ clinical characteristics and behaviors, histologic slides, radiologic images, and so on. For these data, descriptive methods from the subdiscipline of qualitative research are appropriate. For ethical and legal reasons, subjects’ personal information must always be de-identified when qualitative findings are presented.

Similarly, numerical data on individual subjects are almost never presented in a research report. Numerical information is usually presented as statistical summaries such as averages and measures of variability. The mean, median, and mode are all measures of central tendency. The measure of variability most often used in medical literature is standard deviation. Occasionally, a range or standard error (SE, also called standard error of the mean, SEM) is used. When confidential information is not disclosed, numerical reports may be supported by exemplar, anecdotal, or model information that conveys valuable research or teaching lessons.

**Analytic Studies**

Analytic studies test for the following:

- Differences between groups (e.g., a group of patients receiving a drug compared to those receiving a placebo)
• Relationships among variables (e.g., the correlation between cholesterol levels and coronary artery occlusion)
• Both differences and relationships

A “true experiment” is a prospective design in which the researcher controls as many subject, treatment, and environmental variables as possible.

A cross-sectional study is a database of “snapshots” of subjects at one period. For example, childhood bone maturation may be studied by describing bone densities in a group of children ages 1, 3, 6, and 9 years old. A longitudinal approach would follow a group of 25 one-year-old children over the next 8 years. If the researcher could control all or most variables over time, then a longitudinal study could also be experimental. However, it would be difficult to determine whether intervening variables had corrupted the design. Therefore, long-term experiments involving human subjects are rare.

When there is substantial risk to human subjects, it is not always possible to conduct a true experiment ethically. The risk may be caused by inducing the disease or condition being studied or by testing an experimental intervention or treatment. As an alternative, it may be possible to study a disease (e.g., cervical cancer) by identifying those who already have it and those who do not and then comparing the two groups for factors that might have been responsible for the disease (e.g., human papillomavirus exposure). This type of study is known as a case-control study design.

Another experimental approach is to identify people who have or do not have a particular risk factor (e.g., human papillomavirus exposure) and then examine the two groups over time to identify those who develop the disease or condition. This type of study is known as a cohort study design.

Prospective Versus Retrospective Designs

Data collection for prospective studies is planned in advance. Prospective designs include true experiments and concurrent cohort studies. The concurrent cohort design involves subjects who do not have the disease in question but have a suspected risk factor and are tested at a later time to determine whether the disease developed in them. In many cases, the prospective approach is the only way to obtain information about a new or recently discovered phenomenon, such as a previously unrecognized disorder or a newly developed technique.

True experiments attempt to gain strict control over the conditions of the study, including subject selection, instrument calibration, and the experimental environment. Compared to retrospective designs, prospective designs are credited with having better control of variables and a greater possibility of having valid and reliable standardized measurement methods. Disadvantages include cost (which may limit subject numbers) and the difficulty of extrapolating the results of strictly controlled methods to the clinical setting, where a similar level of control is not possible.

Retrospective studies examine data that already exist, including chart reviews and case-control studies. A retrospective approach may be the only ethical way to study the mechanisms of certain interventions. For example, thalidomide was widely prescribed outside the United States for nausea and vomiting during pregnancy, but later it was found to cause developmental defects. Looking back at the unexpected adverse events associated with a previously approved therapy is an example of a retrospective approach. In hindsight, the harms discovered provide valuable information that cannot be ethically studied in a prospective manner.

...whole milk leads to heroin addiction? A control or reference group composed of people who do not use heroin is likely to show many have used marijuana or have consumed whole milk.
Because the researcher may not know exactly how and under what circumstances the data were collected, it may be difficult to verify that retrospective data were collected properly. Collecting data retrospectively is usually inexpensive, and data are readily available in large quantities. A large \( N \) value (the total number of participants in a study) may be used to compensate for variability. Therefore, the value and usefulness of a study should not be based solely on whether the data were collected prospectively or retrospectively. The particular method of data collection should be evaluated on its own merit.

Uses of Statistics

Many skilled clinicians are skeptical or even fearful of statistics. They contend that “anything can be proved or disproved,” and that readers are easily fooled. This is not true for those who have knowledge about research methods and the basic statistical principles presented in this chapter. Statistical information is common in medical literature. A few easy-to-learn concepts enable clinicians to become informed consumers of medical information. This ability is invaluable for interpreting study results in journal articles, conference presentations, and drug advertisements. Because we rely on scientific information throughout our professional careers to understand new information, a healthcare provider should be familiar with clinical research methods and basic statistical terminology.

Statistics are used primarily in three ways in medical literature. First, statistics are used to describe and summarize group information. Second, they allow us to infer or generalize sample results to the larger population, which is essential because it is impossible or impractical to measure each and every individual. Third, statistics test for significant relationships or differences between groups of subjects.

Types of Data

A datum (singular of data) is a unit of information about a subject. An example would be the total cholesterol value of 178 mg/dL on a subject participating in a lipid study. Data may be non-numerical (nominal or ordinal, also referred to as nonparametric because of the statistical tests that can be legitimately performed on these types of data) or numerical (interval or ratio, also called parametric).

Nominal data are characteristic names that have no numerical value; examples include male or female, black or white, osteoarthritis or rheumatoid arthritis, smoker or nonsmoker. Ordinal data have characteristics that are comparative and can be ordered by rank; examples are shorter or taller, less painful or more painful, darker or lighter, or an increased size of a palpable mass. Interval data have numerical values between units but no actual zero point; examples include degrees centigrade and blood glucose (i.e., a patient could not have a body temperature or blood glucose value of zero). Ratio data have numerical values between units, and a zero value is possible; examples include milligrams of alcohol per deciliter, basophils per cubic millimeter, and number of pack-years smoking history (number of packs smoked per day multiplied by the number of years smoked).

Like clinical data, numerical research data are preferred to non-numerical data. For instance, in a study on smokers, knowing the number of pack-years (numerical data) provides better information than knowing only that a patient smokes (non-numerical). The more powerful parametric statistical tests can be applied to numerical data.

Although it is not always possible to quantify data, data can be valuable in the descriptive, qualitative form. For example, a narrative about a radiograph of a hairline fracture with a detailed description of its important characteristics, landmarks, and possible origin may be more informative than just a report of a “hairline fracture: present or not present” on 100 films.

When it is not possible to use numerical data, nonparametric statistical tests (such as the Spearman’s correlation or chi-squared) must be used. The application of parametric tests (e.g., Pearson’s correlation or the \( t \)-test) to nonparametric data is a mistake because a falsely low \( p \) value...
(probability value) is likely to be fabricated (i.e., referred to as an alpha or type I error). A falsely high \( p \) value is likely to result when nonparametric tests are used for analyzing numerical (parametric) data. In this case, it is unlikely that a false hypothesis will be accepted, but consequentially a true one may be overlooked (i.e., a beta or type II error). This may occur especially when the calculated \( p \) value is close to the desired alpha level.

These classifications of numerical data are commonly used in most scientific articles. However, qualitative research methods also serve invaluable roles in some medical studies and are used as extensively in clinical research as in clinical problem solving.

**Experimental Designs**
An investigator must decide whether a qualitative or quantitative approach will more effectively address the research objective. A second decision is whether the study will involve retrospective data (e.g., clinical records) or prospective data (i.e., information that will be newly generated as part of the project). The implications of these decisions have been discussed earlier in this chapter.

If the investigator decides to use an experiment to answer the research question, then choosing the appropriate experimental design is imperative. In a project proposal, the investigator describes the methods to be used for analyzing the study’s data. Many proposal guidelines suggest creating mock data that are likely to resemble the study’s actual data. This process also helps the investigator plan for both computer needs and possibly consultation with a statistician or methodologist.

The selected research designs presented in Table 6–1 are likely to accommodate the needs of most investigators. Although experimental and epidemiological studies share some methodological similarities, epidemiologists have developed specialized techniques for the needs of their discipline. Some investigators may need to use advanced epidemiological techniques for their studies. In these cases, consulting an epidemiology textbook or an epidemiologist is appropriate. The statistical principles presented here apply to most experimental and epidemiological studies, whether prospective or retrospective. Familiarity with these principles should enable the clinician to converse knowledgeably with most clinical research methodologists and consultants.

**Pre-Experimental, Experimental, and Quasi-Experimental Designs for Research**
The purpose of this section is to provide the reader with a simple way to recognize common research designs and the associated threats to validity inherent in those designs. Most of the work conducted in this area can be attributed to Campbell and Stanley, who wrote the classic text *Experimental and Quasi-Experimental Designs for Research*. They developed a shorthand method for diagramming research designs similar to the way English grammar has been classically taught—by diagramming sentences, students learn to understand proper grammar and sentence construction. Similarly, individuals can more easily recognize a research design by diagramming it.

Campbell and Stanley and others have used certain conventions. For the purposes of this section, the following symbols are used:

- An \( R \) represents randomization and indicates a particular group was randomly selected or assigned.
- An \( M \) indicates the groups were matched.
- An \( X \) (with or without a subscript) indicates a treatment.
- An \( X_0 \) indicates no treatment or, in some cases, the control condition.
- An \( O \) indicates a measurement. If multiple measurements are taken, then subscripts, such as \( O_1, O_2, O_3 \), and so forth, may be used.
- A dashed line indicates nonequivalency and denotes the groups may be substantially different.
**PRE-EXPERIMENTAL DESIGNS**

The pre-experimental designs are the weakest of the research designs and are subject to many threats to internal and external validity. They are characterized by the lack of a control group, sensitivity to temporal threats to internal validity, and poor generalizability.

**One-Shot Case Study**

The first pre-experimental design, the one-shot case study, is diagrammed as follows:

\[ X \rightarrow O \]

In this design, a treatment is given and a measurement is made. This particular design is typical of survey research. A group of respondents is identified based on one or more pre-existing criteria and are administered a questionnaire that is then measured. Another example of the one-shot case study is the typical high school or college classroom. The instructor assumes the students have a certain level of baseline knowledge when they enter the class, but this baseline is not measured. The treatment consists of the exposure in the class, and the measurement is the students’ performance in the class. If the students do well, then the instructor may conclude it is because the students learned a great deal from the class. However, the instructor cannot justifiably arrive at that conclusion because the students’ prior level of knowledge in the subject is unknown. This demonstrates how the one-shot case study is a weak design because of its vulnerability to threats to both internal and external validity.

Even with its limitations, the one-shot case study is useful for certain types of research, such as descriptive studies in which the investigator wishes to describe what currently exists. For example, most surveys can be characterized as one-shot case studies. Often the data they yield are very useful to clinicians and, therefore, are published in the medical literature. For example, Levine was interested in the clinical practice of lung transplantation and wanted to ascertain if there were wide differences among transplantation programs. He surveyed 65 active lung transplantation programs. By requesting information on key areas of practice, such as lung preservation and posttransplantation care, he was able to conclude that there was substantial consensus as well as a few areas of variance among transplantation centers. The results could not be generalized to centers that did not respond. In addition, these studies do not have a long “shelf life” because a survey is most often a snapshot of a phenomenon at a specific time. However, the information could be considered valuable because lung transplantation centers may be able to determine if the practices at their center are consistent with those at other centers.

**One Group Pretest–Posttest**

Slightly more robust is the one group pretest–posttest design. In this particular design, subjects receive a pretest, the treatment, and a posttest or retest:

\[ O \rightarrow X \rightarrow O \]

This design is characteristic of clinical practice in which a patient (or a group of patients) is evaluated and diagnosed, a treatment is administered, and the patient (or a group) is then subsequently re-evaluated.

For example, a group of high school students is provided with education about alcohol abuse. The pretest and posttest may be a questionnaire on alcohol use. This design is particularly vulnerable to the temporal threats to internal validity, and therefore the investigator theoretically cannot conclude the outcome was the result of treatment because the illness may have been self-limiting or the subject may simply have outgrown the problem as a result of development. In the case of high school students, changes in alcohol use may be related to the time of year and lack of parties where alcohol is served rather than to alcohol abuse education.

Without the ability to compare the study group to a reference or control group, the effects of temporal threats to internal validity cannot be evaluated. Similarly, the results cannot be generalized...
because it is not known whether the patient (or group) is representative.

In a study on the effects of a 5-day immersion leadership development experience on current and aspiring nursing leaders, Tourangeau et al. assessed and compared self-appraisals by participants and assessments by colleagues before and 3 months after the experience. Although they concluded “a concentrated leadership experience is effective in strengthening leadership behaviors,” critical readers may point out several issues that challenge their conclusions. Because participants’ colleagues assessed leadership behaviors before the experience and knew they would be completing another assessment later, they may have become more sensitive to observing leadership behaviors. In other words, some of the observed differences may be due to changes in the raters rather than in the participants. The study could have addressed the concern about differences in the raters by having them also rate individuals who did not undergo the leadership program. Ideally, the raters would have been blind to the subjects’ participation, or in other words, they would not know whether a person they were rating had or had not participated in the leadership development program. If changes in leadership were observed for individuals who had not participated in the program, then some of the observed change could possibly be attributed to the changes in the raters. This concern about the raters is further supported by the fact that the participants’ self-report on leadership behaviors was not significantly different after the program. Finally, because the subjects were volunteers, those who felt they would benefit from the program may have self-selected to participate. Hence the subjects may not have been representative of aspiring nurse leaders.

Static Group Comparison

The third pre-experimental research design is entitled the static group comparison. In this design, a group that has received a treatment, or been exposed to a condition, is compared to a group that did not receive the treatment or was not exposed to the condition:

\[ \text{XO} \]

\[ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \]

\[ \text{O} \]

The dashed line indicates that these groups are not equivalent. This design, which attempts some form of control with the group that was not treated or exposed, is better than the previous designs. However, because the groups were not randomly assigned, it cannot be concluded that they are equivalent.

This design is commonly used in environmental or occupational epidemiological studies. For example, one may be interested in a group of people who are living in an area that might be contaminated with a carcinogenic agent. The investigator may hypothesize that living in this area may result in a greater incidence of certain types of cancers. People who live in the contaminated area are compared with similar people who live in an uncontaminated area. If there is a greater and statistically significant difference in the number of people with cancer in the contaminated area, then the researcher could conclude the suspected carcinogen might be related. Please note that one cannot conclude causality from this type of design. Correlation and causation are different concepts. Correlation helps to identify a relationship between variables and the direction of that relationship.

The usefulness of the static group comparison is that it allows the study of variables that generally cannot be manipulated by the investigator. Legally and ethically, an investigator cannot require people to live in an area that is suspected to cause a disease or disability. On the other hand, if individuals choose to live in that area, then one can measure and analyze the effects and compare them to a similar group of individuals who live in a different “control” area.

An example of a static group comparison from the medical literature is a study by Lynch et al.
that compared residency choices by medical students who did or did not participate in an optional rural health awareness program. One group of medical students participated in an enrichment initiative entitled the Rural Health Scholars Program (RHSP), which included both didactic and experiential learning focusing on practice in underserved and rural areas. The proportion of participants that chose residencies in primary care, family medicine, community hospitals, and known underserved areas were compared to their classmates who did not participate in the RHSP. Although the results demonstrated that significantly more students who participated in the RHSP chose residencies in family medicine and community hospitals, the authors could not conclude that the RHSP was the “cause.” Because the groups were nonequivalent, the possibility that medical students self-selected participation in the RHSP because of a possible interest in a rural primary care setting could not be ruled out.

In summary, one of the major problems with a static group comparison occurs because the groups are nonequivalent. Therefore, there is always the possibility that some factor other than the treatment is causing the results.

**EXPERIMENTAL DESIGNS**

The next series of research designs are the true experimental designs. These studies are characterized by randomization of the subjects and a control group. Randomization does not mean the groups are identical. Randomization works on the law of probability, which suggests that when a group is selected or subjects are assigned randomly, the traits, characteristics, and conditions that may affect the outcome (confounding variables) are distributed roughly equally among the groups, thus canceling out their effect.

**Pretest, Posttest, Control Group Design**

The first experimental design includes a pretest, posttest, and control group. In this design, subjects are randomly assigned to a group, pretested, given a treatment or not given a treatment, and then tested after exposure to the treatment or no treatment:

\[
\begin{align*}
R & O_1 X_0 O_2 \\
R & O_1 X_1 O_2
\end{align*}
\]

In this diagram, two groups are compared on a single variable with two conditions (\(X_0\) and \(X_1\)), but more than two groups or conditions can be analyzed using this design. In addition, it is possible to examine more than one variable.

The pretest, posttest, control group design effectively rules out most threats to internal and external validity. Because it is the most rigorous design, the pretest, posttest, control group design is commonly accepted as the “gold standard” and is typically used in randomized clinical trials. It should be noted that in clinical trials and other research involving human subjects, it is unethical and often illegal to withhold treatment. Thus the control group may be the group that receives the conventional treatment while the experimental group receives the new treatment. In other words, the conventional treatment is sometimes used as a control or the baseline.

In a study to examine whether fortified cereals actually increased subjects’ blood levels of selected B vitamins, Tucker et al. compared two groups of older adults—one who ate a fortified cereal and one who ate a cereal that was not fortified. Their approach was to measure blood levels of homocysteine (an amino acid), folic acid, vitamin B_6, and vitamin B_12 in subjects on two occasions to obtain baselines. According to the authors, “High homocysteine and low vitamin B concentrations have been linked to the risk of vascular disease, stroke and dementia.”  

After obtaining baseline measures, the subjects were randomly assigned to a group that ate either a cup of fortified cereal daily or a cup of cereal that was not fortified. At 12 and 14 weeks the blood levels of homocysteine, folic acid, B_6, and B_12 were measured.

Neither the subjects nor the investigators performing the tests knew whether a subject was consuming fortified cereal or not. When the group membership is unknown to the subject and the investigator (until
the tests are concluded), the study is called a double-blind study. Both the subject and the investigator are “blind” to the treatment. Comparisons between the two groups revealed significant differences between them. Homocysteine levels were lower, whereas folic acid, $B_6$, and $B_{12}$ levels were higher in the group that ate the fortified cereal. Because of the rigorous design of the study, the authors concluded that eating a fortified cereal benefits older adults. Because major threats to internal and external validity were controlled, the observed differences were due to the treatment. Therefore, the results may be generalized to similar groups of individuals; that is, eating a cereal fortified with B vitamins benefits older adults.

**Posttest-Only Control Group Design**

The second experimental design is termed the posttest-only control group design. The subjects are randomized, given a treatment, and then the results are measured and compared:

$$R \times_0 O$$

$$R \times_1 O$$

Typically, this type of design is used when a pretest is inappropriate or unavailable for other reasons. For example, an orthopedic group may be interested in the outcomes of a total hip replacement when using one particular appliance versus a second appliance. In this case, a pretest is inappropriate because patients who require total hip replacements generally lack substantial range of motion at the hip joint, often because of pain. In this example, the outcomes such as time to ambulation, pain, and range of motion may be appropriate to measure after the surgery. Because it is impossible to know the baseline or the pretest condition of the individuals, there might be some threats to internal validity based on pre-existing conditions. However, randomization minimizes these effects.

Gallo and Staskin investigated the effect of external cues on patient compliance in performing pelvic-floor exercises. Because compliance is measured after a patient is educated about a treatment program, a pretest was not appropriate. In this study, 86 women with stress urinary incontinence were given a program on pelvic-floor exercises and then were randomly assigned to a group that either received an additional audio cassette or did not. They were subsequently evaluated for compliance to the pelvic-floor exercise program, and the results were compared. The women who received the audio cassette, labeled an external cue by the authors, demonstrated significantly better compliance than the group who received only instruction. In this case, the robust experimental design, a posttest-only control group design, controlled major threats to internal and external validity, allowing the authors to conclude that external cues improve compliance for women with stress urinary incontinence.

**Solomon Four-Group Design**

The next experimental design is the Solomon four-group design. In this design, groups with and without pretests are exposed to one of two treatments and are subsequently tested after treatment:

$$R \times_0 O_0$$

$$R \times_1 O_0$$

$$R \times_0 O_2$$

$$R \times_1 O_2$$

The Solomon four-group design is actually a combination of the previous two designs: the pretest, posttest, control group design and the posttest-only control group design. This particular design is useful because it allows an investigator to assess whether an effect occurs because of the pretest. On the other hand, this design requires twice as many subjects.

Danley et al. investigated the effect of a multimedia learning experience designed to prepare dental students and dentists to recognize and respond to domestic violence. The investigators were interested in ensuring that observed changes in the subjects between the pretest and the posttest were not due to the fact that the subjects completed the pretest. When the subjects who completed the pretest, the multimedia learning experience, and the posttest were compared to subjects who completed the multimedia learning experience and the posttest, but no pretest, investigators found
no significant differences between the groups. On the other hand, the two groups who completed the multimedia learning experience were more knowledgeable than the subjects in the control group.6

QUASI-EXPERIMENTAL DESIGNS

The final group of research designs is termed quasi-experimental designs. Although they are more rigorous than the pre-experimental designs, they are not as rigorous or as robust as the true experimental designs. Typically, they lack one or more of the characteristics that would make them true experimental designs. Generally, randomization or multiple measurements are lacking, thereby making testing effects a potential problem.

Nonequivalent Control Group Design

The nonequivalent control group design is similar to the pretest, posttest, control group design, but because the groups are not randomized they must be considered nonequivalent:

\[ O_1 X_0 O_2 \]
\[ O_1 X_1 O_2 \]

An example of the use of this design is a hypothetical comparison of patient satisfaction with care from two clinics: one clinic hired a clinician 6 months previously (X1) and one clinic had no clinician (X0). Many ways in which the patients differ might affect the outcome, such as age, years seen by the provider, and so forth. Because the groups are nonequivalent, the results obtained from studies using this particular design must be viewed with caution; pre-existing conditions in the subjects may account for the changes after treatment.

Kristjansson et al. studied the effectiveness of a short-duration learning experience for teenagers that focused on preventing skin cancer.8 Because the presentation was given in a classroom, they compared the classes that received the presentation to classes that did not. In this example, randomly assigning students would have made the study very complex and unmanageable. Although their results suggested the short-duration learning experience was successful, the fact that the groups were nonequivalent suggests that they should be viewed with caution. For instance, it is not known if one or more classes were composed of students who were more concerned about their health or more motivated to accept the content presented.

Separate Sample Pretest, Posttest Design

The separate sample pretest, posttest design is used when the investigators suspect that the pretest will significantly bias the posttest results:

\[ O X \]
\[ X O \]

Although measuring a change from pretest to posttest is appropriate when using a physical test such as blood level or heart rate, a pretest of a mental, psychological, or performance variable often affects the posttest. For example, a pretest for manual dexterity may result in some motor learning that could affect subsequent performance. If the investigator were interested in evaluating the effect of a new treatment, the observed changes could be due to motor learning from the pretest. The separate sample pretest, posttest design addresses concerns about the pretest biasing the outcome.

Markert et al. used a separate pretest, posttest design to assess the knowledge acquired by professionals who attended continuing medical education (CME) programs.9 The investigators noted in this article that pretesting and posttesting attendees was an inappropriate approach for assessing knowledge. They felt that a pretest would bias performance on a posttest. Given the observation that most medical professionals are bright individuals who are committed to learning, it seems logical that they would remember items from a pretest. Instead, Markert et al. randomly assigned attendees to either a pretest group, which they considered a “control” group, or a posttest-only
group, which they considered the “treatment” group. Upon comparing the groups, the treatment group demonstrated significantly more knowledge in selected topic areas.9

**Time Series Design**

A third quasi-experimental design is the time series design. In these designs, groups are compared to each other and there are multiple tests:

\[
O_1 \times_0 O_2 O_3 O_4
\]

\[
O_1 \times_1 O_2 O_3 O_4
\]

Because there are multiple measurements, one cannot conclude that changes or differences are not a result of sensitization or learning. In other words, the threat to validity of repeated testing is of concern, as well as the threat that groups may be nonequivalent.

In a study that compared nursing and medical interventions for skin care, Pokorny et al. examined differences in patients who developed pressure sores and those who did not.10 Data were gathered on the day of admission, the day of surgery, and the following 4 days. Because the patients were admitted for cardiac surgery, it was anticipated a subject would average 6 days in the hospital. The investigators in this study looked for differences in the daily nursing and medical interventions between patients who developed pressure sores and those who did not. For this study, the dependent variable (what was measured) consisted of the documented interventions, specifically differences. The authors concluded that multiple assessments of skin condition by nurses were important in preventing pressure sores.10 Because patients were not randomly assigned to groups that received standard or optimized skin assessment, differences in patients may have affected the results.

**SUMMARY**

The designs presented are by no means exhaustive of the pre-experimental, experimental, and quasi-experimental designs found in the literature. By recognizing the common designs, one can quickly focus on threats to validity. When designing a study, knowledge of common research designs may help avoid pitfalls. Although one can have a robust research design, failure to obtain reliable measurements using a valid instrument may render the conclusions of the study inaccurate or erroneous. Finally, reading research critically or performing research is a learning experience that generally requires a great deal of practice.

When evaluating the methods section of published research reports, this information should help clinicians to knowledgeably critique the methods used and evaluate the study’s results before applying them to practice. For all clinicians, this information should help in interpreting new medical information presented in literature, professional conferences, and through other sources. Well-informed clinicians may be able to interpret and apply clinical information in new or better ways.

By definition, original research studies are unique. Applying the principles presented in this chapter should help beginning investigators select an effective research method for a planned research study and, when necessary, communicate effectively with methodological consultants.

**REFERENCES**


† The astute reader will undoubtedly observe that the example used is a slight variation from the classical pretest, posttest, control group design because the investigators used two baseline measurements and two posttest measurements, so the design was:

\[
\begin{align*}
R \quad O_1 \quad O_2 \quad X_{\text{not fortified}} \quad O_3 \quad O_4 \\
R \quad O_1 \quad O_2 \quad X_{\text{fortified}} \quad O_3 \quad O_4
\end{align*}
\]

Ordinarily one might also be concerned about pretest–posttest sensitization with multiple tests, but in this study the dependent variable was blood levels of homocysteine, B6, B12, and folic acid, which should not be affected by repeated testing.