2

Designing an Outcomes Research Study

David M. Radosevich

TYPES OF STUDY DESIGNS

A health outcomes research study design is a plan for executing the study. At a minimum, the design depicts the groups studied; for example, treatment and control group, instances of the treatment and the timing, and frequency of health outcomes measures. The design provides a high-level overview of the health outcomes study and insights into the plan for analysis. Finally, the design should specify whether the individuals studied are randomly assigned to either receive the treatment of interest or no treatment, also referred to as a control group.

Control over treatment assignment through randomization is the basis for distinguishing two types of outcomes studies: experiments and quasi-experiments. Random assignment of subjects is central to controlling for extraneous differences between groups, but it does not guarantee comparability; it simply asserts that any differences are due to chance. Without randomization of study participants, the outcomes researcher runs the risk of individuals with particular characteristics having a higher probability of being included in the study or one of the study groups. These differences can arise from patient self-selection or from clinician decisions about who should get treatment. Selection bias or self-selection has the potential to confound the treatment-outcome relationship, thereby biasing results. Some of these differences can be measured and controlled for in the analysis, but others may remain unmeasured and uncorrected. Overall, selection bias may be the greatest threat to the validity of health outcomes research studies.

Although the randomized controlled trial (RCT) is considered the "gold standard" for clinical research, most outcomes studies are conducted as quasiexperiments, which lack control over the assignment of participants to receipt of treatment. As a consequence, the outcomes researcher is faced with controlling for self-selection and underlying differences between treatment and no treatment groups, by other means such as the timing of the outcome measurement (relationship to randomization? can be done in observational studies) or statistical adjustment. Many treatments cannot be practically investigated using an experimental design. In outcomes studies conducted in health plans, fairness is a frequently voiced concern regarding allocating individuals on the basis of randomization (Disease Management Association of America Outcomes Consolidation Steering Committee, 2004). Consequently, the quasiexperimental design, also called the observational study (Kelsey, Whittlemore, Evans, & Thompson, 1996), serves as the backbone of health outcomes research.

Self-Criticism in the Design Process

There is no perfect health outcomes research study. Every investigator must weigh trade-offs between internally valid designs, like the RCTs, and quasiexperiments where inferences could be erroneous because of an inability to randomly assign treatments to study participants. Designing an outcomes research study requires a process of self-criticism and self-evaluation. This is accomplished by raising questions concerning the validity of the study design or the accuracy of inferences drawn. In this iterative process of self-criticism, the outcomes researcher comes to recognize the imperfection of the study design, its strengths and limitations, and identifies strategies for strengthening the overall design. In truth, validity encompasses all the grey areas of research and is always context specific. It is mistaken to interpret the validity of study designs as simply good or bad.

The goal of a health outcomes research study is to achieve the best approximation of the truth of the treatment-outcomes relationship. Does a given treatment cause a particular outcome? Understanding and evaluating the threats to the validity of human inferences about the results of an outcomes study is critical to success. This involves addressing four study design questions. The remainder of this chapter discusses the implications of these questions and the common threats to the validity of health outcomes studies.

EVALUATING THE THREATS TO OUTCOMES RESEARCH

Validity concerns the truth or falsity of propositions about cause. (Cook & Campbell, 1979). Although a discussion of the multiple threats to study designs is outside the scope of this chapter, a selected few, which are frequently encountered in outcomes research study designs, need to be considered in planning and implementation. They are listed in Table 2–1. For a complete discussion of validity and study designs, the reader is referred to the texts by Campbell & Stanley (1963); Cook & Campbell (1979); and Shadish, Cook and Campbell (2002). For a more humorous treatment of validity, also referred to as bias, see the papers by David Sackett (1979) and Alvin Feinstein (Feinstein, Sosin, & Wells, 1985).

Table 2–1 Adaptation of Cook and Campbell's Scheme (1979) for Classifying Threats to the Validity of Health Outcomes Research

Internal Validity

- Statistical Conclusion
  - Low statistical power
  - Fishing and error rate problems
  - Violated assumptions and inappropriate statistical tests
  - Reliability of measures
  - Inconsistent implementation of the intervention
- Internal Validity
  - Selection
  - Regression to the mean
  - Attrition
  - Missing data
  - History

External Threats

- Construct Validity
  - Inadequate conceptual design
  - Monooperation and monomethod biases
  - Treatment diffusion
- External Validity
  - Person
  - Setting
  - Time

Internal Versus External Validity

Under Cook and Campbell's scheme, threats to validity can be classified as either internal or external. This distinction neatly divides threats into those that concern the validity of conclusions drawn about the relationship between the treatment and the outcome and whether the results are externally applicable to other persons, places, and time. Internal validity is the minimum necessary to interpret an outcomes study. All outcomes studies need to be internally valid; that is, the study design avoids errors that could compromise conclusions. For example, the researcher wants to avoid drawing spurious conclusions regarding results because the subjects in the groups being compared are not comparable.

Issues around external validity concern the generalizability or representativeness of study results. Can the results of an outcomes study be applied across different populations of persons, in different settings, and in other periods of time? Generalizability questions usually can be traced to the methods of recruitment of study subjects. RCTs have been criticized for their lack of generalizability, because study conclusions are limited to the population being studied. Recruitment may employ strict inclusion and exclusion criteria for enrollment; therefore, individuals recruited bear little resemblance to individuals seeking health care in the "real world." Many RCTs rely on volunteers, who themselves are highly self-selected. In contrast with RCTs, quasiexperiments have the potential for being more representative.

Even RCTs can encounter selection bias when the rate of follow-up is poor or even worse when it is different in treatment and experimental groups. The standard way to handle such loss is through a process known as intention to treat (ITT). Basically, the last observation is carried forward as the final observation for that subject. Thus, someone who leaves treatment early is retained at the state when they were last observed. This approach is generally conservative for treatments designed to improve the situation, but it can have the opposite effect if the treatment is simply designed to slow the rate of deterioration. Thus, it must be employed thoughtfully.

A second aspect of external validity concerns the validity of inferences drawn about higher order constructs or traits that cannot be directly observed. Can one generalize from the operational definitions used in the study to abstract constructs? From this perspective, external validity concerns the measurements concepts, the interrelationship of the concepts with one another, and integrity of the treatment investigated. This form of validity is referred to as construct validity. There is a theoretical basis for construct validity in two approaches to constructing outcomes measures: latent-trait theory and factor analysis. According to latent-trait theory, the individual's responses to an item on an outcomes measure depend on the level of the attribute present (Streiner & Norman, 1995). Factor analysis, on the other hand, attempts to represent a set of variables as a smaller number of constructs (Kim & Mueller, 1978). Both latent-trait analysis and factor analysis are useful techniques for confirming construct validity.

Four Study Design Questions

The process for evaluating study designs was best articulated in educational psychology by Campbell and Stanley (1963). Their disciples built on this early work, expanding it to include applications in health services research, epidemiology, and clinical research. More recent work (Cook & Campbell, 1979; Shadish, et al., 2002) stressed the importance of four critical questions in the design of scientific experiments. These questions, which reflect four major threats to the validity of outcomes study designs, have been restated to make them relevant to outcomes research.

1. Is there a relationship between the treatment and outcome?
3. Is the observed relationship between treatment and outcome causal?
4. What concepts explain the treatment outcome relationship?
5. How representative is the treatment and outcome relationship across persons, settings, and times?

Each question relates to a form of validity: statistical conclusion, internal, construct, and external validity respectively. The process of designing a health outcomes research study involves its critique and redesign.

Epidemiologists describe the threats to validity as biases or systematic errors in the design or implementation of a study (Szklo & Nieto, 2000). In addition to confounding, biases are categorized according to type: selection or information. Selection bias was mentioned earlier. Information biases are unique to epidemiological studies and concern misclassification of the treatment (or exposure), the outcome, and imprecise definition of study variables. Two types of misclassification errors are recognized in clinical studies. First, First, one could falsely conclude that the individual received the treatment of interest or has the outcome of interest. These are referred to as false positives. The second type of misclassification error is one in which the individual did not receive the treatment or the outcome of interest. These latter are referred to as false negatives. Together, these errors define the accuracy of treatment and outcomes measures, thereby defining the sensitivity and specificity of measures (Szklo & Nieto, 2000). In applying the internal and external validity scheme discussed earlier, information biases bridge both forms of validity. Epidemiologists have recognized that misclassification significantly biases the results of observational studies. For a full appreciation of how these biases impact outcomes studies, several sources offer a fundamental review of epidemiological methods (Kelsey et al., 1996; Szklo & Nieto, 2000). What follows is a brief review of threats to validity that are important in health outcomes research.

STATISTICAL CONCLUSION VALIDITY

Question 1: Is there a relationship between the treatment and the outcome? This statistical question concerns whether the treatment and the outcome covary and the strength of association between them. Five threats to statistical validity are commonly observed in outcomes research study designs. These include low statistical power, fishing and error rate problems, violated assumptions of statistical tests, reliability of the outcome measures, and inconsistent implementation of the intervention.

Low Statistical Power

All too frequently, the first question asked by researchers is: how many subjects do I need for my study? This question is always premature before planning the study and preparing an analysis strategy. Planning for statistical power begins by addressing the following questions:

- What is the research question?
- Who is the target population?
- How am I going to recruit study subjects?
- How large an effect is expected?
- How much variation is anticipated?

- What analysis is needed to answer the question?
- It is feasible to study the question?

All health outcomes studies need to be designed to detect differences between persons receiving the treatment and those not receiving the treatment. The goal is to detect a true effect. Formally stated, the primary concern of statistical power is the likelihood of detecting the truth about the treatment-outcome relationship.

The determinants of sample size can be best understood in the context of hypothesis testing. For example, in a study to investigate the difference between the risk of occurrence of adverse outcomes between a "new" medical treatment e and usual care signified by *c*, one sets up a hypothetical test scenario as follows.

Null Hypothesis: $P_e = P_c$

Alternative Hypothesis: $P_e \neq P_c$

Where P_e represents the probability of the event among experimental subjects and P_c the probability of the event among controls

Statistics test the likelihood that an observed difference occurred by chance. Where In a study is designed to test for differences in the adverse event rates between the "new" treatment and usual care, the determinant of the sample size is statistical significance level, also called the type I error rate (α); it reflects the likelihood that one sees a difference that could simply have occurred by chance. This is equivalent to the risk of drawing a false conclusion that there is a difference between P_e and P_c. By contrast, a type II error claims no difference when in fact one exists. Statistical power, or one minus the type II error (β), is the probability of discovering a true difference between P_e and P_c. Next, the size of difference considered important is considered. The latter is defined in terms of the effect size, a standardized difference, which reflects how large a difference one wants to be able to demonstrate statistically. Finally, one considers the number of subjects or the number of groups necessary. Examining the interrelationship of the type I error rate, the type II error rate, and the magnitude of the effect being sought is referred to as statistical power analysis.

Many factors under the direct control of the outcomes researcher directly affect the statistical power of outcomes studies. Figure 2–1 shows the impact of various threats to validity on sample size. In the center of the figure, there is general function for estimating sample size (Friedman,

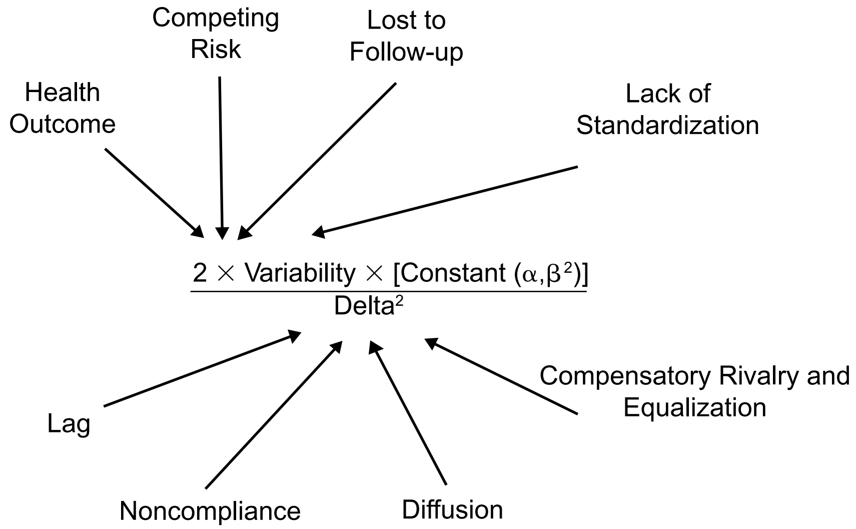

Figure 2–1 Threats to Validity and How They Impact on the Statistical Power of a Health Outcomes Study

Furberg, & DeMets, 1996). Delays in implementing a treatment intervention (lag), individuals not taking prescribed study medications (noncompliance), and treatments spilling over to individuals assigned to the control group (diffusion, compensatory rivalry, and equalization), all increase the number of individuals needed to detect differences between treatment and control groups. In the numerator, poor standardization of the study, individuals being lost to follow-up (attrition), persons dying from causes other than target condition (competing risk), and selecting a health outcome measure with poor responsiveness characteristics inflate sample size. Low statistical power is a recurring threat to the validity of health outcomes research study designs. The best advice is to seek the consul of an expert, preferably someone who will conduct the analysis.

Planning and implementing a health outcome study is a collaborative endeavor. Because statistical power is critical to the design and planning an outcomes study, statistical power should always be specified in advance through an a priori power analysis. Using the results from published studies and knowledge regarding the outcomes measure, it is possible to make an "educated guess" regarding the likely size of the effect of the intervention. This is quantified in terms of an effect size or detectable difference. It may be asked another way: How big a difference is needed to convince the target audience that the treatment effect is meaningful? Although this enhances the efficiency of study designs and eliminates frivolous outcomes studies, most statistical power analysis is done at the end of the study. This post hoc power analysis is only justified under circumstances in which the investigator lacks advanced knowledge regarding the size of treatment differences. Post hoc power analysis should always be done where no statistically significant differences were found in the analysis to be sure a real difference has not been overlooked. The sample size needed to support a claim of no difference is usually much larger than that needed to show a difference.

The following are some guidelines for statistical power analysis.

- Consult with an expert; remember that statistical power analysis and estimating sample size is a collaborative endeavor.
- Sample size should be specified in advance.
- Set standards for considering statistical power.

    1. Type I error less than 5%
    3. Type II error less than 20%
    4. Lowest common denominator for comparison
    5. Plan for available data; response rates, eligibility, missing data

- Be guided by research objectives; consider monetary and indirect costs.
- "Parameters on which sample size is based should be evaluated as part of interim monitoring." (Neaton, 2001)
- Be conservative in estimates of statistical power.

Design a study for the smallest subgroup analysis planned and available data. Guided by the study objectives, always consider the monetary and nonmonetary costs of a study and be conservative in estimates. It is generally a good idea to continuously evaluate the study assumptions as the research unfolds.

Most analysts perform statistical power analysis using any one or a combination of three resources. Formulas for the direct calculation of statistical power can be found in a number of different sources (Friedman, et al., 1996; Murray, 1998; Schlesselman, 1982). In general, most formulas incorporate measures of the type I error, type II error, and the effect size. The arithmetic expression in the center of Figure 2–1 provides a conceptual depiction of the interrelationship of these elements. Tables for frequently used statistical tests can be found in resources such as Cohen (1988). Finally, shareware over the Internet and commercial available software, such as nQuery Advisor (Elashoff, et al., 2000) are available.

Statistical power analysis calculations are more involved when it comes to multivariate analysis. Because these frequently involve complex computations and multiple design parameters, these are best left to skilled biostatisticians, epidemiologists, and health services researchers.

Fishing and Error Rate Problems

Most intuitively recognize that if a researcher conducted a hundred tests of statistical significance, 5 percent would be statistically significant (at the 5 percent level) by chance. Yet frequently outcomes studies are designed, making multiple comparisons and ignoring chance in interpreting the statistical tests. These are collectively referred to as “fishing and error rate” problems. The inflation of type I errors is particularly troublesome in outcomes studies, especially in studies using multiple outcomes and multidimensional scales.

This threat to validity arises when investigators fail to specify their end points in advance of conducting their study or the primary outcomes are ill defined. In the absence of specifying primary end points, the investigator incorporates multiple outcomes measures in their study. When analyzing their results, each of the outcomes is treated as having primary importance to answering the study question, thereby converting the study to one that is more exploratory than hypothesis driven.

A second threat involves the use of multidimensional measures; for example, a new treatment is hypothesized to improve the quality of life of participants. The investigator chooses a multidimensional scale to measure quality of life, such as the Medical Outcomes Study 36-Item Short Form Health Survey (SF-36) (Ware, 1991; Ware & Sherbourne, 1992; Ware, Snow, Kosinski, & Gandek, 1993) or the Sickness Impact Profile (Bergner, 1989; Bergner, Bobbitt, Carter, & Gilson, 1981). Without specifying a specific subscale from these measures, the investigator increases the likelihood of type I errors by treating all the subscales as equally important to the confirming the hypothesis.

Various methods have been devised to adjust for an inflated type I error rate (Rosner, 1995; Shadish, et al., 2002):

1. the LSD approach
2. Bonferroni correction
3. Tukey's method
4. Scheffé's method.

A thorough discussion of these techniques can be found in most biostatistics (Rosner, 1995) and epidemiologic (Last, 2001) texts. In general, the approaches adjust the type I error rate downward, making it more difficult to reject the null hypothesis (no difference) and thereby reducing spurious associations.

Strategies for minimizing error rate problems in health outcomes research studies include:

- Recognize the problem of making multiple comparisons
- Establish a priori the primary outcomes for the study
- Incorporate greater specificity in outcomes measures
- Make adjustments for multiple comparisons selecting one of the accepted statistical techniques, such as Tukey or Scheffé

Violated Assumptions of Statistical Tests

This threat to statistical conclusion validity involves selecting an inappropriate statistical test to answer the study question and violating the assumptions for the statistical tests being used. Although a discussion of the full range of statistical tests applicable to health outcomes research is beyond the scope of this chapter, some general guidelines should be kept in mind. The nature of the variables affects what statistical tests should be used. Tests differ for categorical or a continuous variables and for nominal, ordinal, interval, or ratio scales. (See Chapter 4– for a discussion of scales.)

A variety of techniques could be used for analyzing outcomes. For categorical outcomes, such as death, morbidity, and hospitalization, the rigorous assumptions of normally distributed errors can be relaxed. In some instances, categorical data may be desirable because it allows the researcher to contrast elements that are critical to understanding the effects of a treatment. For example, whereas a linear model that uses mean age may not get at the effects of age 85+, a model that compares those 85+ to those younger might address the issue more directly. In the biomedical literature, logistic regression is widely used to analyze categorical outcomes, such as death or morbidity (Allison, 2001; Hosmer & Lemeshow, 2002; Kleinbaum & Klein, 2002; Le, 1998). Based on a logit function, the technique can be used to simultaneously adjust for covariates and expanded to ordered categorical end points called ordered logit. Multivariate logistic regression is frequently applied techniques for analyzing categorical outcomes data in biomedical studies. Logistic regression yields the odds ratio as the measure of association between the treatment of interest and the outcome (Szklo & Nieto, 2000). A variant, multinomial logit can be used when there is more than one outcome of interest.

Analysis of categorical outcomes can be further strengthened if the investigator knows about the timing of occurrence of the outcome. In this case, time-to-event analysis, also called survival analysis, has been widely used (Allison, 1995; Hosmer & Lemeshow, 1999; Kleinbaum, 1996; Le, 1997). Survival analysis is a powerful technique to analyze time-to-event data. In general, survival analysis improves statistical power for analyzing categorical outcomes by using the time to occurrence of an event to weight the importance of that event. However, this may be easier said than done. The timing of some outcome events may be hard to determine precisely or impossible to obtain. An outcomes investigator might procure death records to determine the date of death for a subject or administrative data to ascertain the date of hospitalization, but the onset of an acute myocardial infarction (MI) can be clouded by an uncertain history of previous MIs, misdiagnosis, and a failure to seek medical care needed to record the event., Finally, one rarely knows the specific date of onset of some conditions such as disability or chronic diseases such as diabetes mellitus.

A second analytic issue is best illustrated by the use of general linear regression in the analysis of continuous outcomes such as health status scale scores, blood pressures, and laboratory values. Replicate outcomes, such as baseline and follow-up health status measure, have the potential to be correlated. The correlated nature of these repeated measures makes it unjustified to use traditional fixed-effects models. Using general linear regression fails to account for the correlated nature of the outcomes measure, thereby artificially increasing the type I error rate for the statistical test (Murray, 1998). In recent years, mixed-model methods have become widely used to handle correlated data (Liang & Zeger, 1993; Littell, Milliken, Stroup, &

Wolfinger, 1996; Murray, 1998). This analytic method has two applications in outcomes studies. Mixed model methods are used in the analysis of data from studies with repeated outcomes measures. This is equivalent to repeated measures analysis and is the basic design for the pretest/posttest study. Outcome study participants are repeatedly measured (e.g., serial blood pressure readings), completing a health status survey at regular intervals. This approach is important in outcomes studies because replicate measures are highly correlated. General linear regression fails to take this into account.

Mixed-model methods are appropriate when the units of assignment or sampling include factors other than the individual, a frequently encountered problem in health services, and public health research. Public health services are frequently delivered at the level of the community. Individuals within communities are more similar than individuals outside the community. The community as a source of variability, called a random effect, is nicely handled using mixed-model methods. By extension, mixed effects can be hospitals, schools, clinics, health plan, or any other type of grouping. These mixed-model methods can be applied to categorical or continuous outcomes.

Some sources of random error beyond the individual participant occur within study designs that draw participants from other units of interest, such as hospitals, clinics, health plans, and communities. These are often referred to as hierarchical or nested outcomes designs in which the treatment or intervention may be influenced by the level of the group—hospital, clinics, and so on. These designs lend themselves to mixed-model methods because of the correlated nature of their data within the level of the group (Murray, 1998). These are called mixed models because there two or more random effects, random effects at the level of the participant, and random effects at the level of the group. Table 2–2 summarizes the appropriate types of analytic approaches for different combinations of distributions, random effects, and data types. Table 2–3 gives recommended guidelines for analyzing outcomes study data.

Reliability of Outcomes Measures

The failure to reliably measure either the treatment or the outcome could result in a misclassification of the treatment status, the outcome, or both. The reliability of a measure imposes an upper bound on that measure's ability to discriminate. Unreliable measures can attenuate the relationship between the treatment status and the outcome, mask a relationship, or create a spurious relationship. This underscores the importance of the routine

Table 2–2 Classification Scheme for Statistical Approaches Useful in Analyzing Health Outcomes Data

	Distribution	
Design Characteristics	*Normal Distribution*	*Nonnormal Distribution*
One Random Effect	General Linear Model—Ordinary Least Squares Linear Regression	Generalized Linear Model—Logistic Regression
Two or More Random Effects/Replicate Outcomes Measures	General Linear Mixed Model	Generalized Linear Mixed Model—Nonlinear Mixed Models
Time-to-Event	Survival Analysis—Kaplan-Meier Life Table Methods and Cox Proportional Hazards Regression	

measurement of reliability of study measures and implementing corrective steps to increase reliability. The following actions can be taken to improve reliability (Shadish, et al., 2002): increase the number of measurements (i.e., increasing the number of raters or items) and choose better quality measures (i.e., better raters and better training of raters; quality items).

Inconsistent Implementation of the Intervention

One of the more serious threats to the validity of outcomes studies in field settings is the consistent implementation of the intervention. Treatment implementation is notoriously unreliable in observational studies. In natural settings, the investigator rarely has control over the treatment implementation. In community settings, treatments frequently lack standardization and are often idiosyncratic to the settings in which they occur. Epidemiologists have long recognized that treatments implemented inconsistently can lead to spurious results (Szklo & Nieto, 2000). If the implementation of the intervention lacks standardization, results are more likely to suggest that there is no treatment effect. Although this is classified as a statistical threat, remedies focus on tight quality control of the treatment implementation and careful monitoring of the implementation. Systematic training of study subjects and staff involved in treatment implementation is critical. This involves the use of implementation manuals, the development and implementation of programs for training staff and subjects, and continuous reinforcement of expectations in order to improve adherence.

Table 2–3 Recommended Guidelines for Analyzing Health Outcomes Study Data

1. What is the nature of the study outcome?
   a. Categorical outcomes can be either nominal, dichotomous, or ordered categorical.
   b. Continuous variables are in a raw form or require a transformation, e.g., cost data is highly skewed and should log transformed.
   c. Time-to-event analysis

2. Is the data highly correlated? Aside from random errors within the subject, are there different sources of random error?
   a. Correlated outcomes data and data with multiple sources of error are best handled using some form of mixed-model method. Linear mixed and nonlinear mixed model methods are useful along with generalized estimating equation approaches.
   b. Noncorrelated data might use a general linear model or logistic regression.

3. Adopt generally acceptable standards for statistical power.
   Type I error less than 5%
   Type II error less than 20%
   Lowest common denominator for comparison
   Plan for available data: response rates, eligibility, missing data

In general, it is easier to measure outcomes than it is to measure treatments. Outcome studies should measure treatment. This is accomplished by monitoring whether a standard treatment was delivered, received by the subject and adhered to. The processes of delivery, receipt and adherence should be incorporated into all outcomes studies. If the researcher has measured the treatment, it is possible to compare the outcomes for those receiving varying levels of the treatment. However, it is possible that subjects self-selecting levels of treatment. Hence, this is generally viewed as weak evidence for an outcomes effect. It is better to use these data to supplement the preferred analytic method, intent-to-treat (Shadish, et al., 2002).

OTHER THREATS TO INTERNAL VALIDITY

Question 2: Does the observed relationship likely reflect causation from the treatment to the outcome or might this same relationship reflect the effects of other factors? This distinction concerns the validity of inferences drawn about the observed relationship. This concern falls into what has previously been described as threats to internal validity. Shadish, Cook, and Campbell (2002) describe nine threats to internal validity that might bias inferences drawn about the treatment outcome relationship. Five of these threats hold particular relevance to health outcomes research study designs: selection, regression to the mean, attrition, missing data, and history.

Selection

Selection is the most serious internal validity threat in health outcomes research studies. As mentioned earlier, selection occurs because at the beginning of the study, on average, individuals in the treatment group differ from those in the nontreatment group in both known and unmeasurable ways. This difference frequently occurs because the treatment cannot be randomly assigned to individuals. Selection is a major problem in case-control studies (Schlesselman, 1982) in which the investigator has difficulty finding a comparable control group for cases that are of study interest. If the study involves hospitalized patients, exposure and risk increase the likelihood of hospitalization, leading to a higher rate of exposure among hospitalized cases than hospitalized controls. The observed distortion in the relationship is referred to as Berkson's Bias (Berkson, 1946; Last, 2001).

The treatment–outcome relationship could be confounded by differences between the treatment and control group. For example in epidemiological studies of obesity and all-cause mortality, the relationship is confounded by cigarette smoking. Smokers often have a leaner body mass but are at increased risk of sudden death, cardiovascular disease, and cancer from cigarette smoking. One approach to deal with the problems of selection is the use of propensity scores (Rosenbaum, 2002). Logistic regression is used to predict membership in either the treatment or control group. Propensity scores derived from the logistic regression are used to match subjects, thereby minimizing group differences across study variables. However, propensity scores cannot account for unmeasured variables that may be the source of the selection bias.

Selection can interact with other threats to internal validity, such as history, attrition, and regression. The following are examples of these interactions:

- Selection-history interaction: An outside event affects one of the groups more than another.

- Selection-attrition interaction: One group of participants are more likely than another to withdraw or drop out from the program.
- Selection-regression interaction: This is a problem of differential regression. In other words, one of the groups by being sicker or healthier is more likely to be average at a later date.

Regression to the Mean

Some outcomes studies are designed by selecting individuals on the basis of being very sick or healthy. For example, in orthopedic surgery studies, one selects study subjects on the basis of those having the poorest functioning and in need of a joint replacement. Using the same functional status measure before surgery and after surgery, individuals typically look "average" after surgery and hence appear to have improved. This tendency to obtain scores approaching the average with remeasurement is called regression to the mean.

Because all outcomes measures carry some level of uncertainty or error in their administration, outcomes measures are never perfectly reliable. The lack of reliability in outcomes measures exaggerates regression to the mean; that is, an unreliable measure is more prone to regress to the mean with replicate administration. In order to minimize the risk of regression, do not select comparison groups on the basis of extreme scores and use measures with demonstrated reliability. Poor reliability in an outcomes measure can be obviated by avoiding single-item indexes and employing multi-item scales.

Attrition

The bane of most outcomes studies is attrition, also referred to as experimental mortality. Study participants fail to complete the outcomes measure that is administered or they drop out of the study. The more frequently an outcomes measure is planned for collection, the greater the possibility that there will be attrition. This is a special type of selection that occurs because subjects drop out after the study begins or certain data is missing. Using the earlier orthopedic surgery example, following surgery, individuals fail to return for follow-up and hence do not complete the planned-for outcomes measures. Randomization of subjects fails to control for the effects of attrition. Individuals with poorer results from the treatment or those with less education might be less likely to return for follow-up or complete study measures. This selective attrition biases results across groups making results applicable to those that are better educated and most benefited by the treatment.

A related facet of attrition that is a form of selection is survival selection. This occurs when there is a correlation between patient survival and the treatment studied. For example, in observational studies involving patients with AIDs, those surviving longer are more likely to receive the treatment (Glesby & Hoover, 1996). When treated and untreated patients are compared, the treated group appears to have a longer survival. Survival bias can also distort results. If only survivors are compared, the group with the better survival rate may appear worse because the most vulnerable died. One way to counter this effect is to include those who died in the assessment of outcomes; for example, death may be treated as the worst functional state.

Missing Data

In outcomes studies, data will always be missing. The best way to minimize threats posed by missing data is good quality control. This includes careful study management, well-defined project protocols, and clear and well-thought-out operations. Continuous monitoring for quality control minimizes missing data. In addition, it is always best to use available data rather than discarding study variables or cases. Missing data is positive information.

Murphy's Law for outcomes research could read: "If there are any ways in which data can be missing, they will be" (Cohen & Cohen, 1983). Observations needed for conducting outcomes research could be missing for a number of reasons. Attrition, which was discussed earlier, is one reason for missing data. In health outcomes questionnaires, individuals may skip questions either accidentally or deliberately. In other cases, information requested might be difficult or impossible for participant to provide (e.g., questions are too personal or difficult), data systems crash and cannot be recovered, or measuring instruments, such as automatic blood pressure machines, fail. Missing data threatens the integrity of outcomes research and greatly complicates statistical analysis. It threatens the validity of statistical conclusions drawn, particularly if the method for handling missing data is unacceptable and introduces systematic bias. Missing data effectively reduces data for analysis by attenuating statistical power; thereby, reducing the likelihood of detecting differences.

The best solution for missing outcomes is improved quality control in the data-collection process. In order to effectively reduce problems posed by missing data, it is critical to distinguish between the types of missing data (Cohen & Cohen, 1983). Is the data missing for the outcome or the treatment? If the outcome is missing, the investigator is faced with dropping the subject from the study. This can lead to a comparison of unbalanced groups, less-representative samples, and a loss of statistical power.

Is data randomly or selectively missing? Health survey researchers expect a certain amount of random nonresponses in every study. If the pattern of nonresponse is equally distributed across all subjects, it should not introduce a systematic bias. Selectively missing data poses a more serious problem. Selectively missing data is frequently encountered in studies of special populations, such as the elderly or persons with mental health problems. In studies of the elderly, those with cognitive deficits are less likely to provide risk factor data than those with full cognitive function (Radosevich, 1993), but individuals who are unable to provide requested information about their baseline status are at higher risk for poor health outcomes.

Are many versus few items missing? As a general rule, no more than 1 to 2 percent of values should be missing for outcomes study variables. If the pattern of missing values shows that certain data is missing more frequently, then questionnaires and data collection forms should be revised.

Dropping variables with high rates of missing values may be safer than dropping subjects (Cohen & Cohen, 1983). Some investigators elect to drop variables from their analysis if extensive data is missing. If the data being dropped makes no material contribution to the outcomes study, dropping it is of little consequence. In that case, the investigator might reconsider why the variable was included in the study. Resources were wasted and information is being lost.

On occasion, the investigator chooses to drop participants from the study. In many advanced statistical packages used for analyzing health outcomes data, this procedure is referred to as listwise deletion. If the data is an outcome, as noted earlier, dropping participants might be perfectly justified. However, beyond 1 or 2 percent of participants, this could introduce significant attrition bias into studies. The outcomes study overall loses statistical power and becomes less representative of the target population. This selective loss of subjects is an unacceptable strategy for handling missing data.

Pairwise deletion of participants is generally found in studies using correlation methods or bivariate techniques. Associations are examined only for the paired observations in which the study factor of interest and outcome are both present. If data is randomly missing, this approach might work, but the investigator is unclear as to the study population.

A number of acceptable methods for handling missing data in outcomes studies have been suggested. First, use dummy variable codes for missing values. This means that in the place of a missing value for a variable, one employs a dummy code that flags the variable as missing. In the analysis, this strategy has the effect of creating an additional variable for the missing factor and quantifying potential bias introduced by the absence of a value for that variable.

A second group of techniques involves the interpolation of outcomes values: (1) carrying the last observed outcome forward, (2) interpolation between known outcomes, and (3) assuming the worst outcome. Different assumptions underlie each of these approaches. For example, in a study of the long-term follow-up of mechanical heart valve recipients, individuals lost to follow-up are assumed to have died from a heart value–related condition. This involves assuming the worst case scenario. As an alternative, one might assume the individual was still alive because that was their status at the time of their last contact.

Finally, mean substitution is an extension of linear regression techniques and frequently used where the outcome variable is derived from a multi-item scale. The basis for this approach is that the best predictor of missing value is the other values for the same individual. For multi-item scales such as the 10-item Physical Functioning Scale (PF-10) score (McHorney, Kosinski, & Ware, 1994; Ware, 1991; Ware & Sherbourne, 1992; Ware, et al., 1993), a mean scale score is computed on the basis of available items. If the individual completes 7 of the 10 items comprising the PF-10, the scale score is based on the available seven items. This approach underscores an additional advantage of using multi-item scales.

History

History concerns events that occur between the treatment and the outcome that are outside the control of the researcher. For example, in an observational study of the effectiveness of primary care–based treatment program for diabetes mellitus, the introduction of a new drug to treat diabetes is likely to affect the outcome. In the real world, it is impossible to isolate routine care from external changes that occur in the health care environment.

Because it is impossible to control for outside events, outcomes researchers have employed several strategies to control for history effects. First, investigators have attempted to isolate the study participants from the outside. This can more easily be accomplished in the laboratory than in the field. In laboratory experiments, study participants receive the experimental intervention in a setting isolated from the field. In field settings, assignment groups could be separated from one another; for example, hospitals and clinics located in different communities. A second strategy to reduce history effects is to use nonreactive outcomes measures. Examples of these measures include laboratory tests and physiological measures that are less susceptible to outside effects. A third strategy is to use a control group drawn from a comparable group of participants. If the intervention and the control groups are comparable and outcomes measurements occur at same time, history effects would be uniform across the study groups. They might minimize the effect of treatment but will not inflate it.

CONSTRUCT VALIDITY

Question 3: What constructs are involved in the particular cause-and-effect relationship? Constructs are abstractions inferred from specific observations. Outcomes research is generally concerned with drawing conclusions about attributes that cannot be directly observed. For example, one cannot directly observe physical functioning in a subject but can observe the manifestations of physical functioning (e.g., walking, climbing stairs, standing from a seated position). Physical functioning is a construct.

Construct validity involves understanding of how the concepts used in the model relate to one another and how they are measured. There are a number of threats to construct validity (Shadish, et al., 2002).

Inadequate Conceptual Design

Inadequate preoperational explication of constructs simply means that the measures and treatments have not been adequately defined and explained prior to implementing the study. Many investigators fail to adequately define and analyze the concepts they are studying. Before commencing an outcomes study, an operating model needs to be spelled out. At the very least, the following conceptual planning needs to occur (see Chapter 1):

- Develop and define concepts in the operating model.
- Create an operational model of the study that shows the relationship between concepts.
- Operationally define and critique the concepts.

Good conceptual planning is at least as important as choosing the right measures.

Monooperation and Monomethod Bias

Monooperation concerns using only a single measure of each concept; for example, using a single outcomes measure or treatment measure. Single operations of study constructs pose the risk of not measuring the concept in the correct way or measuring irrelevant facets of the concept. One way of reducing this threat is by using a number of instances of the concept. For example, to reduce this bias in the measurement of treatment, the design could incorporate various doses of an experimental drug. This strategy would enable the investigator to demonstrate a dose-response relationship. Alternatively, the investigator might increase the number and type of interventions. For example, in a diabetes mellitus disease management program, different forms of patient coaching might be employed: nurse telephone calls, patient learning materials, and physician coaching.

Monomethod bias is a related threat to validity, wherein a single method is used to collect data. For example, a study of the effectiveness of a particular diabetes intervention might use only self-reported survey data to answer the question. This design is susceptible to a monomethod bias threat. It would be better to include other measures of effectiveness such as laboratory values or medical records review. The distinction between monooperation and monomethod bias is often not clear. For this reason, they might be lumped as the monobias threats. They include what measures are used to assess the concepts and what data collection methods are employed.

Treatment Diffusion

Treatment diffusion is a recurring problem in observational studies. Participants in the control group sometimes receive some of the treatment.

This is sometimes called a "spillover effect." For example, in a study of the effects of a hospital nursing care model, hospital units not receiving the intervention could be exposed to the intervention through staff contact between units implementing the care model and those not implementing the care model. Unknown to the investigator, nursing units assigned to the control condition could implement facets of the nursing care model. The diffusion of the treatment would be likely to attenuate differences in the outcomes between the treatment and control conditions. At a minimum, one needs to look for the possibility of a diffusion effect. Other designs can mitigate this effect—for example, by allocating treatment to different physicians or clinics—but this design imposes other problems.

EXTERNAL VALIDITY

Question 4: How representative is the relationship across persons, settings, and times? Randomized controlled trials are the backbone of biomedical research and the "gold standard" for determining the efficacy of medical therapies. Through randomization of participants to treatment conditions, the RCT gains in terms of internal validity. However, a major limitation of the RCT is the lack of generalizability. Because RCTs use strict criteria for the inclusion and exclusion of study subjects, results are not as representative of the persons and settings of greatest interest in health care and outcomes research. Moreover, RCTs are costly to implement.

Although observational studies suffer from many of the threats to internal validity discussed earlier, they more successfully represent the populations that are receiving the care. The representativeness, also called generalizability, applies to three facets of study designs: the individuals participating in the study, where the treatment occurs, and the timing or time interval for the study. No single design can adequately address the threats to validity. There are tradeoffs. The most discussed and major tradeoff is between internal and external validity. Some argue that timely, representative, and less-rigorous observational studies are to be preferred over internally valid study designs. There are no hard and fast rules.

Table 2–4 summarizes the threats to validity discussed. These are a few of many possible threats but are the ones that hold greatest relevance to health outcomes research studies. For those listed, the table briefly defines the threat, provides an underlying cause, and gives some possible solutions.

Table 2–4 Adaptation of Cook and Campbell's Scheme (1979) for Classifying Threats to the Validity of Health Outcomes Research

Validity Threat	*Definition*	*Underlying Cause*	*Possible Solution*
Statistical Conclusion Validity			
Low statistical power	Study design does not permit detecting a true effect	Inadequate sample size and responsiveness of outcomes measure	Increase sample size; choose an outcomes measure with optimal responsiveness characteristics
Fishing and error rate problems	Multiple comparisons increase the likelihood of making a type 1 error	Too many hypotheses; lack of a primary hypothesis	Identify primary and secondary hypotheses; post hoc adjustments for making multiple comparisons
Violated assumptions of statistical tests and inappropriate statistical test	Inappropriately applied statistical test or the assumption of the statistical test is violated	Careless analysis; plan for analysis not well thought out; failure to consult with an analytic expert	Consult with an analytic expert; use a statistical method that takes into account the correlated nature outcomes data
Reliability of measures	Unreliable outcome measures	Selecting unstable measures; lack of standardization of measurement	Monitor the quality of measurement; select measures based on sound psychometric properties
and treatment implementation	Inconsistent implementation of the treatment	Lack of standardization of treatment implementation; lack of clarity regarding treatment implementation	Monitor the quality of treatment implementation; take corrective measures to assure standardization; closely monitor the treatment implementation; incorporate treatment measures into study design

Table 2–4 Adaptation of Cook and Campbell's Scheme (1979) for Classifying Threats to the Validity of Health Outcomes Research

Validity Threat	*Definition*	*Underlying Cause*	*Possible Solution*
Internal Validity			
Selection	Differential selection of subjects to the treatment and control groups	Failure to randomize treatment to subject groups	Risk adjustment; propensity analysis
Regression to the mean	Selection of sicker or healthier subjects for the study more likely to result in outcomes at follow-up that look average	Recruitment criteria for the study focuses on sicker or healthier subjects; unreliable outcome measures	Use of a control group with similar characteristics to the treatment group; improve the reliability of the outcome measure
Attrition and missing data	Subjects drop out or leave the study before its completion	Inadequate follow-up leaves subjects lost to follow-up; death from a cause unrelated	Quality control of the data collection process
History	Events that occur during the study that affect treatment implementation and outcomes	Changes in routine treatment (e.g., introduction of a new medication, changes in reimbursement, patient management) that could have an effect on the outcome	Monitor and document external factors that could affect treatment implementation and outcomes
Construct Validity			
Inadequate explication of constructs	Study concepts are poorly defined and their interrelationship not well spelled out	Failure to develop an operating model for the study; muddled thinking about the question	Adequate planning of the outcomes study with focus on measures and their interrelationship

continues

Table 2–4 Adaptation of Cook and Campbell's Scheme (1979) for Classifying Threats to the Validity of Health Outcomes Research

Validity Threat	*Definition*	*Underlying Cause*	*Possible Solution*
Monomethod and monooperation biases	Using single methods to collect data; using a single measure of the treatment and outcome	Cost prohibitive to use multiple measures of treatment and outcomes; using single methods of data collection	Employ multiple methods to collect the factors of study interest, e.g., written surveys, personal interviews, and physiological testing; employ multiple approaches in measuring the treatment and outcomes, e.g., self-report, interview, observation
Treatment diffusion	In natural settings, the treatment spills over to groups not intended to receive the intervention	Inadequate segregation of treatment and control group subjects; rivalry between groups not given the treatment and those receiving the treatment	Whenever possible, plan to give the control subjects the treatment after the study has concluded; blind subjects to the treatment; give control subjects a "sham" therapy
External Validity			
Representative ness to person, setting, and time	Results of the study limited by person, setting, and time	Inclusion and exclusion criteria limit the findings	Replicate studies across different populations, in diverse setting, and at other points of time

QUASIEXPERIMENTAL DESIGNS

Types of Quasiexperimental Designs

Most outcome studies will be observational and hence will rely on quasiexperimental designs. A bare-bones experimental design has an intervention and an outcome. In the absence of a control group, these are sometimes referred to as preexperiments. The preexperimental design could be expanded by adding control groups and pretest measures. All outcomes study designs can be described using a standard nomenclature (see Table 2–5). The preexperimental design consisting of an intervention and outcome could be depicted as follows:

$$X \qquad O$$

In this posttest-only design, an outcome (*O*) is observed only after a treatment (*X*). This type of design is frequently used in medical practice and is referred to as a case study; patients receive a treatment and the researchers observe an outcome. A number of problems are associated with this design, selection, history, and attrition, to name a few. From a statistical perspective, this design is not interpretable. One cannot observe covariation because the design fails to incorporate either a pretest or a control

Table 2–5 Standard Nomenclature for Describing Quasixperiments

O — outcomes measures or an observation

X — treatment

~~*X*~~ — removed treatment

R — random assignment of subjects/groups to separate treatments

NR — no random assignment of subjects/groups to separate treatments

Subjects/groups separated by dashes – – – – are not equated by random assignment

Subject/groups divided by a vertical dashed line ¦ are not necessarily equivalent to one another group. The simple randomized controlled trial (RCT) adds both a pretest and a control group. The RCT could be depicted as follows:

$$\begin{array}{cccc} R & O_1 & X & O_2 \\ R & O_1 & & O_2 \end{array}$$

In this design, participants are randomly assigned (R) to treatment and control conditions. A preintervention observation (O_1) is made before the treatment (X) is delivered, followed by a postintervention observation (O_2). If the outcome is a measure of physical functioning, it makes intuitive sense to have a measure of functioning before an intervention (e.g., joint replacement surgery or an exercise program). Statistically, this enables the researcher to observe covariation, a necessary prerequisite for statistical conclusion validity. However, in practice, many investigators omit the O_1 measures and rely on randomization to produce equivalent groups.

For irreversible end points, a posttest-only design with randomization would be essentially equivalent to a randomized control trial. For example, in study where survival was the primary outcome, it makes little sense to think about a preintervention measure. The status of the participant is alive at the time of their recruitment. Preintervention observations might include measures of comorbidity and severity for risk adjustment; but randomization, if performed correctly, assures that treatment and control groups are comparable at the time the intervention condition is delivered. Nonetheless, it may prove valuable to collect baseline characteristics to use in the analyses.

If the investigator lacks control over allocating participants to the treatment or control conditions, then the study is described as a quasiexperimental. Using the standard nomenclature, a quasiexperiment investigating the effectiveness of a disease management program might look like the following:

$$\begin{array}{ccc} O_1 & X & O_2 \\ \hdashline O_1 & & O_2 \end{array}$$

In this design, the dashed line is used to signify that the groups are not randomized. From the outset, selection is a serious internal threat to validity. In order to draw valid inferences about differences in the outcomes between the two groups, investigators would need to be able to statistically adjust for differences between treatment and control groups. Any conclusions will be confounded by morbidity differences between the groups.

While a discussion of all possible outcomes study designs is beyond the scope of this chapter, a few design characteristics are worth noting, using the scheme of Shadish and his colleagues (2002): designs without control groups, designs without pretest measures, and combination designs.

Designs Without Control Groups

The posttest-only design described earlier can be improved by adding a pretest measure. This type of approach has been used in evaluating the effectiveness of programs such as disease management and health education (Linden, Adams, & Roberts, 2003).

$$O_1 \quad X \quad O_2$$

The investigator makes a pretreatment observation and looks for a change in that measure with follow-up. Although this provides some evidence for change that could be a result of the intervention, it fails to rule out other things that might have happened to the participants (history), such as other treatments, practice changes, or statistical regression to the mean. One improvement to this design is to add additional pretest measures.

$$O_1 \quad O_2 \quad X \quad O_3$$

Adding multiple pretest measures reduces the threat of statistical regression, but one cannot rule out the possibility that other external factors might have led to the changes that occurred. This type of design lends itself to situations in which repeated pretest measures are available to the investigator. For example, in a study intended to reduce the use of medical services, prior use of services might serve as pretest measures. The lack of a trend in the use of health care services before the intervention strengthens the argument for the effect of the intervention minimizing threats of regression or age. Shadish and his colleagues (2002) discuss other types of designs without control groups such as the removed-treatment design, repeated-treatment design, and designs that use nonequivalent observations. The interested reader is referred to this source for a more comprehensive discussion.

Designs Without Pretest Measures

The pretest is an observation taken before the intervention condition in order to ascertain the preliminary status of the participant. In many outcomes research studies, it is impossible to obtain a pretest measure; for example, participants are emergent cases. In this type of study, a nonequivalent control group might be used. One of the more frequently used design without a pretest is the posttest-only design with nonequivalent groups.

$$\begin{array}{ccc} NR & X & O_1 \\ \hdashline NR & & O_2 \end{array}$$

Here the dashed horizontal line indicates that the group receiving the intervention is different from the control group. Campbell and Stanley (1963) called this the static group comparison. Participants receiving the treatment are compared to those who did not, thereby establishing the effect of the intervention. Certainly, the biggest problem with this type of design is selection; participants in one group could systematically differ from those in the other group leading the observations made. One approach to dealing with this threat is to add an independent pretest sample.

$$\begin{array}{cc:cc} NR & O_1 & X & O_2 \\ \hdashline NR & O_1 & & O_2 \end{array}$$

Here the vertical dashed line signifies that the participants at time 1 and time 2 may be different. These observations are independent of one another. This design is used frequently in epidemiology and public health where it is impossible to collect data on the same group of participants pretest and posttest. The level of intervention is at a group or system level and participant level of control is less critical to the study question. For example, what are the effects of community level intervention to increase smoking cessation?

Combination Designs

Some quasiexperimental designs use both pretests and control groups. The simplest design is the nonequivalent treatment and control group design with a dependent pretest and posttest.

$$\begin{array}{cccc} NR & O_1 & X & O_2 \\ \hdashline NR & O_1 & & O_2 \end{array}$$

The pretest and posttest make it simpler to evaluate the study for attrition (i.e., drop out of subjects) and regression to the mean. However, because participants are not randomized to treatment conditions, differential selection remains a problem; for example, participants receiving the treatment condition are sicker and heavier users of health care services than those receiving the control condition.

One way to improve this design is to add pretest measures or switch interventions.

$$\begin{array}{lllll} NR & O_1 & O_2 & X & O_3 \\ \hline NR & O_1 & O_2 & & O_3 \end{array}$$

This type of design might be beneficial where there are ethical concerns about withholding a therapy that could be beneficial to the participant or demoralizing to participants in the control condition.

$$\begin{array}{llllll} NR & O_1 & X & O_2 & & O_3 \\ \hline NR & O_1 & & O_2 & X & O_3 \end{array}$$

This brief list of typical designs that have been used in health outcomes research is not exhaustive but merely represents some of the more commonly found designs and some thought about how these might be improved.

GENERAL GUIDELINES FOR DESIGNING A HEALTH OUTCOMES RESEARCH STUDY

Evaluate the Threats to Validity

This chapter has identified the threats to validity that are frequently encountered in outcomes research studies. For more exhaustive and comprehensive, the reader is encouraged to explore some of the references cited. The most complete treatment of threats to validity can be found in the works of Cook and Campbell (1979) and Shadesh and colleagues (2002). These authors have built on the earlier work of Campbell and Stanley (1963), establishing the nomenclature for classifying and describing study designs and characterizing biases found in observational studies. The reading is a bit turgid, but worth the effort to gain an appreciation of the multiple layers of quasi-experiments.

The most important message to remember is the need to identify all potential threats to the validity as one is planning a study. Because quasi-experiments are especially susceptible to internal validity threats—including selection, mortality, and statistical regression, much of the effort is focused in this area. The outcomes researcher needs to engage in a continuous process of self-criticism, preferably involving input from peers with expertise in the area of study. Present a proposal formally to colleagues for their review. Although this can be a particularly humbling experience, even for those viewed as experts in their field, the finished product will be much improved.

Construct and statistical conclusion validity are frequently ignored from the outset of design. Investigators will embark on a study before sufficient work has been done developing and refining an operating or conceptual model for their work. As discussed in Chapter 1, this oversight frequently leads to poor operationalization of study variables, ignoring and omitting key factors from their study, and a muddled analysis plan. The conceptual work and statistical plan needs to be undertaken before the beginning of the study. Dealing with these threats is no less important than the work of coming up with a sound internally valid study design. If a researcher can visualize what the final product will look like, it is advisable not to start.

Draw a Study Logistical Plan

When protocols are developed from randomized controlled trials, the study investigator frequently develops a flow diagram called a schedule of events, which demarcates the timing of measurements for the clinical trial. This schedule of data collection provides direction for the study manager about what data needs to be collected, when the data is collected, and from whom the data is collected. The schedule includes all the variables collected as part of the study, the study subjects personal characteristics, their risk factor profile, data necessary for risk adjustment (e.g., comorbidity, disease severity), and outcomes measures such as laboratory values, health outcomes questionnaires, and adverse events. Importantly, the schedule of events includes when the study data is collected and from which study subjects this data is collected.

It is likewise helpful to diagram the overall design of the study. For outcomes research studies, the graphical design demarcates study groups, time dimensions, outcomes variables, multidimensionality, and possible contrasts for analysis.

Graphics can also be an important adjunct in presenting the results of the study. The graphical image provides greater depth and dimensionality that is impossible to communicate verbally. For an excellent discussion of the graphical display of quantitative information, the reader is encouraged to review works by Tufte (1990, 1997, 2001).

Use Design and Statistical Controls

Statistical control, or risk-adjustment control, can never overcome the effects of a poorly designed study. In general, the best strategy is to use a combination of sound study design and statistical controls in implementing the health outcomes research study. Shadish and colleagues (2002) refer to this as the "primacy of control by design." Design involves adding control groups, making additional observations before the treatment intervention, and timing data collection.

Analyzing outcomes data requires statistical techniques that are often beyond the skills of most investigators. The use of correlated methods, described as mixed-model methods, and time-to-event analysis, called survival analysis, requires advanced statistical course work. Because of the complexity of analysis, sound study design must involve input from a skilled data analyst at an early stage of the planning process. This assures that the study question has been clarified, the analysis plan fits the study design, the right variables are being collected, and the study can produce the desired results.

REFERENCES

Allison, P.D. (2001). *Logistic regression using the SAS® System: Theory and application*. Cary, NC: SAS Institute Inc.

Allison, P.D. (1995). *Survival analysis using the SAS® System: A practical guide.* Cary, NC: SAS Institute Inc.

Bergner, M. (1989). Quality of life, health status, and clinical research. *Medical Care, 27*(3, Supplement), S148–S156.

Bergner, M.B., Bobbitt, R.A., Carter, W.B., & Gilson, B.S. (1981). The sickness impact profile: Development and final revision of a health status measure. *Medical Care, 19*, 787–805.

Berkson, J. (1946). Limitations of the fourfold table analysis to hospital data. *Biometrics Bulletin, 2*, 47–53.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally and Company.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin Company.

Disease Management Association of America Outcomes Consolidation Steering Committee. (2004). *Disease management: Program evaluation guide* (1st ed.). Washington, DC: Disease Management Association of America.

Elashoff, J.D., Oliver, M.R., Yeghiazarian, K., Zheng, M., Jamshidian, M., & Koyfman, I. (2000). nQuery Advisor (Version 4.0). Los Angeles, CA.

Feinstein, A.R., Sosin, D.M., & Wells, C.K. (1985). The Will Rogers phenomenon. *The New England Journal of Medicine, 312*(25), 1604–1608.

Friedman, L.M., Furberg, C.D., & DeMets, D.L. (1996). *Fundamentals of clinical trials.* St. Louis, MO: Mosby.

Glesby, M.J. & Hoover, D.R. (1996). Survivor treatment selection bias in observational studies. *Annals of Internal Medicine, 124*(11), 999–1005.

Hosmer, D.W., & Lemeshow, S. (2002). *Applied logistic regression* (2nd ed.). New York: John Wiley and Sons, Inc.

Hosmer, D.W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data.* New York: Wiley-Interscience.

Kelsey, J.L., Whittlemore, A.S., Evans, A.S., & Thompson, W.D. (1996). *Methods in observational epidemiology* (Vol. 26). New York: Oxford University Press.

Kim, J.-O., & Mueller, C.W. (1978). *Introduction to factor analysis: What it is and how to do it* (Vol. 13). Beverly Hills, CA: Sage Publications.

Kleinbaum, D.G. (1996). *Survival analysis: A self-learning text.* New York: Springer-Verlag.

Kleinbaum, D.G., & Klein, M. (2002). *Logistic regression: A self-learning text.* New York: Springer-Verlag.

Last, J.M. (Ed.). (2001). *A dictionary of epidemiology* (4th ed.). New York: Oxford University Press.

Le, C.T. (1998). *Applied categorical data analysis.* New York: John Wiley and Sons, Inc.

Le, C.T. (1997). *Applied survival analysis.* New York: Wiley-Interscience.

Liang, K.-Y., & Zeger, S.L. (1993). Regression analysis for correlated data. *Annual Review of Public Health, 14*, 43–68.

Linden, A., Adams, J.L., & Roberts, N. (2003). An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management, 6*(2), 93–102.

Littell, R.C., Milliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). *System for mixed models.* Cary, NC: SAS Institute, Inc.

McHorney, C.A., Kosinski, M., & Ware, J.E. (1994). Comparisons of the costs and quality of norms for the SF-36 survey collected by mail versus telephone interview: Results from a national survey. *Medical Care, 32*(6), 551–567.

Murray, D.M. (1998). *Design and analysis of group-randomized trials* (Vol. 27). New York: Oxford University Press.

Neaton, J. (2001). Design and conduct of clinical trails. In D.M. Radosevich (Ed.). Minneapolis.

Radosevich, D.M. (1993). *Factors associated with disability in the elderly.* University of Minnesota.

Rosenbaum, P.R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosner, B. (1995). *Fundamentals of biostatistics* (4th ed.). Belmont, CA: Duxbury Press.

Sackett, D.L. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32*, 51–63.

Schlesselman, J.J. (1982). *Case-control studies: Design, conduct, analysis*. New York: Oxford University Press.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Streiner, D.L., & Norman, G.R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). Oxford: Oxford University Press.

Szklo, M., & Nieto, E.J. (2000). *Epidemiology: Beyond the basics.* Gaithersburg, MD: Aspen Publishers.

Tufte, E.R. (2001). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E.R. (1997). *Visual explanations: Images and quantities, evidence and narrative.* Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Ware, J.E. (1991). Conceptualizing and measuring generic health outcomes. *Cancer, 67*(3), 774–779.

Ware, J.E., & Sherbourne, C.D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). *Medical Care, 30*(6), 473–483.

Ware, J.E., Snow, K., K, Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey: Manual and interpretation guide.* Boston: The Health Institute, New England Medical Center.