# CHAPTER 3

# Study Designs and Their Outcomes

*"Natural selection is a mechanism for generating an exceedingly high degree of improbability."*
—Sir Ronald Aylmer Fisher

*Peter Wludyka*

**OBJECTIVES**

- Define research design, research study, and research protocol.
- Identify the major features of a research study.
- Identify the four types of designs discussed in this chapter.
- Describe nonexperimental designs, including cohort, case-control, and cross-sectional studies.
- Describe the types of epidemiological parameters that can be estimated with exposed cohort, case-control, and cross-sectional studies along with the role, appropriateness, and interpretation of relative risk and odds ratios in the context of design choice.
- Define true experimental design and describe its role in assessing cause-and-effect relationships along with definitions of and discussion of the role of internal and external validity in evaluating designs.
- Describe commonly used experimental designs, including randomized controlled trials (RCTs), after-only (post-test only) designs, the Solomon four-group design, crossover designs, and factorial designs.
- Define quasi-experimental design and compare it with true experimental design with respect to validity and assessing causal relationships.
- Describe commonly used quasi-experimental designs, including nonequivalent control group design, after-only nonequivalent control group design, and single-group designs.
- Describe repeated measures designs and how they might naturally be used.

## INTRODUCTION

Most of science can be described as efforts to describe some process or phenomenon or as efforts to discover relationships among entities in the physical world. Particular goal-oriented scientific activities are often called studies. This chapter is concerned with study design, which is used interchangeably with research design or, when appropriate, experimental design. The design arises from (or is structured to fit) the objectives of the study, which is a set of activities undertaken to answer some research

question. The idea is rather primitive and hard to define precisely but is typically an attempt to gain understanding about some process or phenomenon. "Research design provides the glue that holds the research project together. A design is used to structure the research, to show how all of the major parts of the research project—the samples or groups, measures, treatments or programs, and methods of assignment—work together to try to address the central research questions" (Trochim, 2006).

There are many different types of studies with many different purposes. The studies we are interested in concern disease (used very broadly) and its relationship to exposure (again used very broadly). Most studies we are interested in have the following properties:

- A study is conducted to answer some research question.
- The studies we are interested in involve observation (recording, measuring, and such).
- Often, a rather general statement of the research question is turned into a precise statement.
- The research question usually involves the "measurement" of one or more outcomes/responses. Generically, one may denote this as $y$, which can stand for a single outcome measurement (one variable) or several (more than one variable).
- Typically, other variables are "measured" either at the same time or before measurements of $y$. Let's use $x$ to stand for these. Note that $x$ might be something as simple as a dichotomous variable that identifies treatment subjects and control subjects.
- Usually, the researcher is interested in the relationship between $x$ and $y$.

Studies can be described by their features such as randomization, type and degree of control, blinding, timing, whether manipulation or an intervention takes place, as well as the nature of the associated statistical analysis plan.

The study protocol is a detailed description of what will be done and how it will be accomplished. A statistical (data) analysis plan usually accompanies a protocol. For a particular study the protocol and the embedded data analysis plan may be quite detailed. In this chapter, and consistent with usual practice, a design is a more abstract notion and the phrase "study design" is typically used to describe a study with respect to its major aspects in a manner that is often independent of a specific study. That is, design is often a shorthand description of a study that is generic in that many very different studies with very different purposes can be described as having the same design. This approach allows certain commonly used designs to have names and also allows for discussion of strengths and weaknesses of particular designs (or aspects of

designs) often with respect to validity (a subject with many forms and aspects discussed later is this chapter).

The design taxonomy in this chapter divides designs into nonexperimental, experimental, quasi-experimental designs, and time series (repeated-measures) designs. A rather detailed list and description of experimental and quasi-experimental designs appears in Garson (2010). A discussion of nonexperimental designs appears in Rothman (2002).

## NONEXPERIMENTAL DESIGNS

Nonexperimental designs are missing some or all of the features associated with experimental designs, namely

- Manipulation or an intervention
- Randomization
- Control

More will be said about this later in this chapter. Nonexperimental designs can lead to useful information, and certain conclusions can be drawn from data arising from these designs. Causality is difficult to demonstrate with these designs; that is, typically one has to settle for statements regarding association. Importantly, there are circumstances in which experimental designs are not appropriate (typically for ethical reasons) or impossible. Long-term studies that span several years or decades are difficult to arrange as experimental studies.

### Cohort

In epidemiology a cohort is any group of individuals sharing a common characteristic and observed over time. There is no randomization and no intervention (manipulation) has occurred. Measuring disease within one or more cohorts is the goal of cohort analysis. Characteristic(s) defining the cohort may be ethnic, exposure, geographical, or almost anything that creates a grouping of value in studying disease distribution and occurrence. This is seldom as easy as it appears, especially for large cohorts. Typical complications (questions to be addressed) are as follows:

- Who should be followed?
- What counts as a disease occurrence?
- How are incidence (rates and risks) and prevalence measured?
- What constitutes exposure?

The diagram in **Figure 3-1** displays the structure of a cohort design. In this general case the defined population is independent of exposure. That is, selection took place before any of its members became exposed or before their exposures were identified. In a study of this type many exposures can be studied simultaneously. In addition, for many of the subjects in the study precedence relationships can be established (e.g., did exposure to tobacco occur before the occurrence of chronic obstructive pulmonary disease?).

A cohort study can be retrospective (e.g., the study covers 20 years: the researchers begin in 2010, but the subjects in the study are studied beginning in 1990). This type of study might consist of chart data, public records, or whatever sources of data are available. A cohort study may be concurrent (prospective); for example, it may commence in 2010 and consist of newborn babies in a rural hospital with the plan of studying the subjects until they are 12 years old. If this cohort consists of babies born only in 2010, then it is a closed cohort. If the researchers plan to continue to "enroll" babies over the course of the study, then the cohort is open. Several famous cohort studies were geographically defined (e.g., the still ongoing Framingham Heart Study [Shindler, 2010]). Cohort studies can be retrospective, concurrent, or a combination of these. In cohort studies there is often interest in what can be called the "cohort effect." In this context one might be thinking of a complete cohort study as consisting of cohorts as defined typically by the period during which subjects were born, such as decades.

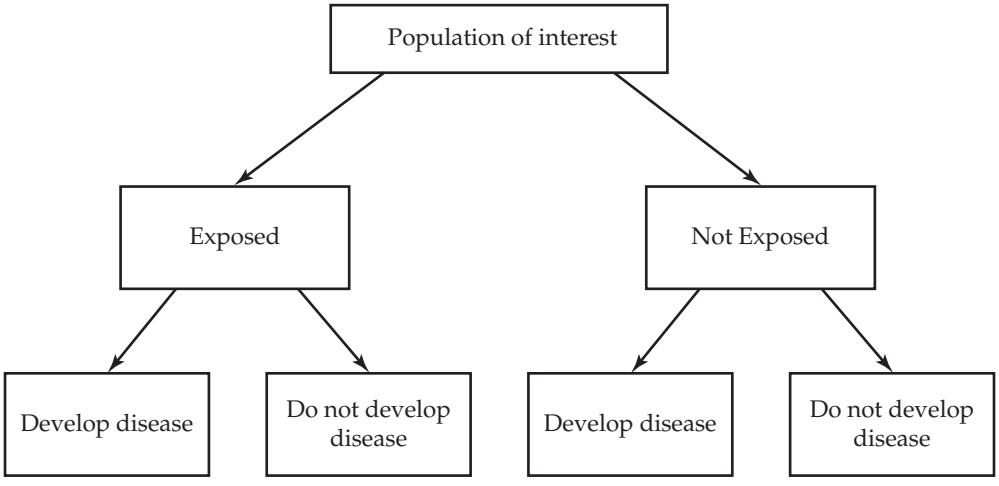


**FIGURE 3-1** Structure of a cohort design.

## Exposed Cohort

An exposed cohort is one in which exposure is the basis for selection (**Figure 3-2**). For example, exposure might be "parent smokes" and the disease of interest might be "child develops asthma." In this type design one might simultaneously study several diseases (or other attributes of the children of smokers). Furthermore, data may be collected on demographics, actual smoking behavior (e.g., "always smoke outside the house"), or other variables of interest. In a cohort study of this type, the subjects in the study are observed over time. Deciding how and at what particular times observation is made are critical issues. Loss to follow-up can also be a critical issue.

Data of any level of measurement (see Chapter 2) can be recorded and analyzed. In epidemiological studies counts are frequently of interest. **Table 3-1** shows how counts data from an exposed cohort study can be arranged in a 2×2 table in which the rows are exposure versus no exposure and the columns are disease versus no disease (in place of disease one could have any well-defined event, such as for example "death prior to age 50"). Note that $a$, $b$, $c$, and $d$ represent counts in which $a$ is the number of those who were exposed and in whom the disease appeared. In a table of this type $n = a + b + c + d$ = the total number of subjects. The key is that in an exposed cohort study the researcher determines the number exposed ($a + b$) and the number with no exposure ($c + d$). For example, suppose that 30 subjects with exposure were selected and 170 with no exposure were selected. What can be estimated from the 2×2 embedded in Table 3-1 that is identified as sample 1 (in which $n = 200$)? One can clearly estimate proportions from the data present (that is, incidence or prevalence proportions). If one wanted to estimate rates, then a different measure of exposure would be needed. For now let's focus on proportions.

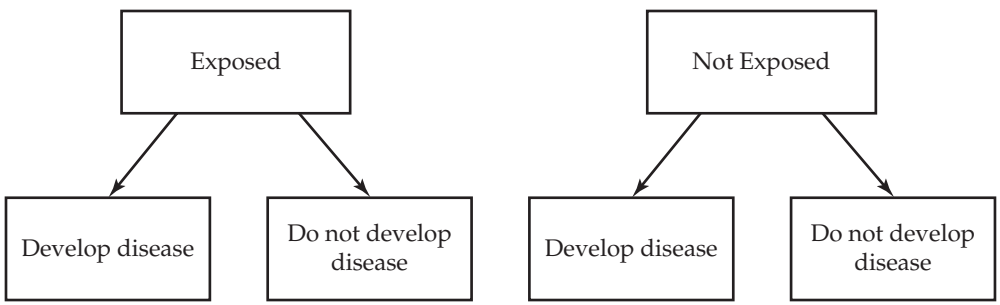


**FIGURE 3-2** Exposed and nonexposed groups.

**Table 3-1 Calculations**

**Observed Counts and Proportions**

| | | Disease | No Disease | sum | prop |
|---|---|---|---|---|---|
| Exposure | | *a* | *b* | $(a + b)$ | $a/(a + b)$ |
| No exposure | | *c* | *d* | $(c + d)$ | $c/(c + d)$ |
| | sum | $a + c$ | $b + d$ | n | |
| | prop | $a/(a + c)$ | $b/(b + d)$ | | |

**Sample Size 1**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 20 | 10 | 30 | **0.67** |
| No exposure | | 80 | 90 | 170 | **0.47** |
| | sum | 100 | 100 | 200 | |
| | prop | 0.2 | 0.1 | | |

**Sample Size 2**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 20 | 10 | 30 | **0.67** |
| No exposure | | 160 | 180 | 340 | **0.47** |
| | sum | 180 | 190 | 370 | |
| | prop | 0.11 | 0.04 | | |

Without loss of generality in the example, the proportions are referred to as prevalence. Consider the following hypothetical study. A nurse researcher selects 30 children about whom it is known that a parent smokes and 170 who do not have a parent who smokes. The children are selected from a pediatric clinic in an urban hospital. Over the course of 1 year episodes of colds or respiratory infection are recorded for the children, and each child is classified as having had an episode defined in this manner (disease) or not (no disease). Data from the study are shown in Table 3-1 (under sample size 1). The researcher can estimate the prevalence of disease given exposure, which is $a/(a + b) = 20/30 = 0.67$, as well as the prevalence given no exposure (0.47). Hence, the relative risk can be estimated, which is $0.67/0.47 = 1.42$ (see more about this subsequently). But the researcher cannot estimate the prevalence of exposure given disease, $a/(a + c)$, because just increasing the number of subjects in the no expo-

sure group (which is 170 in sample size 1 and 340 in sample size 2) can alter the estimate from 20/100 = 0.2 to 20/180 = 0.11. Note that the estimate for disease given exposure is invariant (because the difference between sample 1 and sample 2 is that in sample 2 the number with no exposure has been doubled, keeping the risk of disease fixed). Typically, this would be summed up by saying that, based on data from an exposed cohort study, prevalence can be estimated. That means that parameters derived from risk estimates can be estimated, including attributable risk (see Table 2-19 and the section **Analysis of 2×2 Tables** in Chapter 2, which includes a definition and discussion of attributable risk in subsection Effect Measures Including Attributable Risk and Etiological Fraction) as well as population attributable risk. As a caution, one should not interpret Table 3-1 to mean that, with the proportionate increase in exposure or nonexposure, sample values would be identical. The exercise is strictly theoretical.

## Case-Control

In a case-control study the researcher selects cases and a corresponding set of subjects as control subjects (for a thorough discussion see Schlesselman, 1982). The manner in which the control subjects are selected is important with respect to interpretation and method of analysis. There is no generally agreed-on method for selecting control subjects, and the purposes of the study impact the choice mechanism. Broadly, there is group matching and forms of case matching (with one or more control subjects matched to each case). **Figure 3-3** displays the choice mechanism. The key point is that cases are selected and then exposure is determined; this is performed similarly for control subjects.

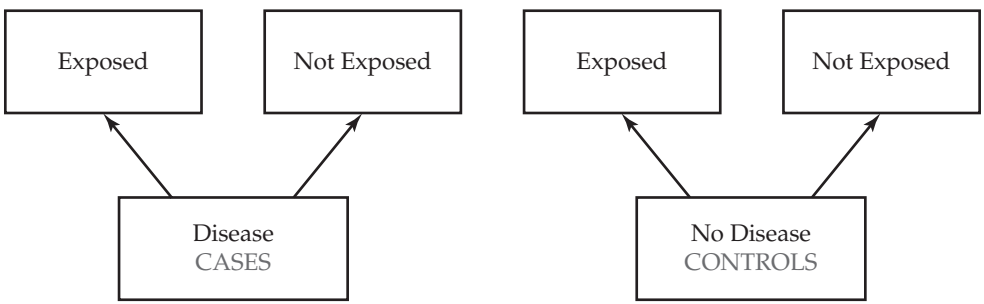


**FIGURE 3-3** Case-control study design.

**Table 3-2** shows the arrangement of a case-control study into a 2×2 table. The definitions are identical to those that apply to Table 3-1. The key idea is that the column sums are fixed (i.e., chosen by the researcher) because the researcher chooses the number of cases and the number of control subjects. The difference between sample 1 and sample 2 is that sample 2 has five times as many control subjects (but the underlying distribution of the control subjects between those exposed and those with no exposure is the same for both samples).

Fixing the column totals implies that one can estimate the prevalence of exposure given disease, which is $a/(a + c)$. Similarly, one can estimate the prevalence of exposure given no disease. The flip side is that one cannot estimate prevalence of disease given exposure, which is $a/(a + b)$, because just increasing the number of subjects in the no disease group can alter the estimate from 0.67 in sample 1 to 0.29 in sample 2. Hence, one cannot directly estimate relative risk. That means parameters derived from risk

**Table 3-2 Formulas and Calculations**

**Case-Control Design: a + c fixed; b + d fixed**

| | | Disease | No Disease | row sum | prop |
|---|---|---|---|---|---|
| Exposure | | a | b | $(a + b)$ | $a/(a + b)$ |
| No exposure | | c | d | $(c + d)$ | $c/(c + d)$ |
| | sum | $a + c$ | $b + d$ | n | |
| | prop | $a/(a + c)$ | $b/(b + d)$ | | |

**Sample Size 1**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 20 | 10 | 30 | **0.67** |
| No exposure | | 80 | 90 | 170 | **0.47** |
| | sum | 100 | 100 | 200 | |
| | prop | 0.2 | 0.1 | | |

**Sample Size 2**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 20 | 50 | 70 | **0.29** |
| No exposure | | 80 | 450 | 530 | **0.15** |
| | sum | 100 | 500 | 600 | |
| | prop | 0.2 | 0.1 | | |

estimates, including attributable risk, cannot be estimated, directly from case-control data; that is, data/estimates from other sources must be found. An alternative measure is the odds ratio. This can be estimated with case-control data and will be discussed subsequently in the Odds Ratios section.

## Cross-Sectional

A cross-sectional study is a snapshot based on a sample of size *n*—that is, the numbers of diseased or exposed subjects is determined by chance within the context of characteristics of the population. The schematic in **Figure 3-4** displays how it works. The key idea is data on exposure and disease are gathered simultaneously. Hence, all that can be concluded from such data is association. However, because neither the row sums (exposures) nor column sums (disease counts) are fixed (predetermined by the researcher) one can estimate both the prevalence of disease given exposure and the prevalence of exposure given disease (**Table 3-3**). The difference between sample 1 and sample 2 is *n* has been doubled. The cell counts are increased proportionately. The fact that risk of disease given exposure can be estimated means that parameters derived from risk estimates can be estimated, including attributable risk as well as population attributable risk.

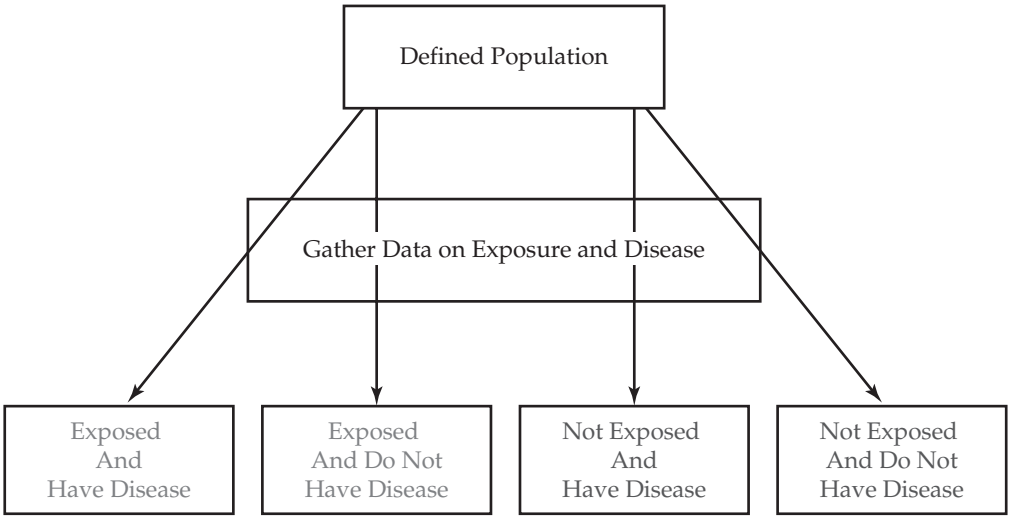


**FIGURE 3-4** Cross-sectional study.

**Table 3-3 Cross-Sectional Study**

**Cross-Sectional Study: $n = a + b + c + d$ is fixed**

| | | Disease | No Disease | row sum | prop |
|---|---|---|---|---|---|
| Exposure | | $a$ | $b$ | $(a + b)$ | $a/(a + b)$ |
| No exposure | | $c$ | d | $(c + d)$ | $c/(c + d)$ |
| | sum | $a$ + c | $b + d$ | $n$ | |
| | prop | $a/(a + c)$ | $b/(b + d)$ | | |

**Sample Size 1**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 20 | 10 | 30 | 0.67 |
| No exposure | | 80 | 90 | 170 | 0.47 |
| | sum | 100 | 100 | 200 | |
| | prop | 0.2 | 0.1 | | |

**Sample Size 2**

| | | Disease | No Disease | row sum | |
|---|---|---|---|---|---|
| Exposure | | 40 | 20 | 60 | 0.67 |
| No exposure | | 160 | 180 | 340 | 0.47 |
| | sum | 200 | 200 | 400 | |
| | prop | 0.2 | 0.1 | | |

## Odds Ratios

Whether the data arise from an exposed cohort study, a case-control study, or a cross-sectional study, the association between exposure and disease can be measured using an odds ratio, and in each case the calculations are the same. Consider first the exposed cohort (Table 3-1). The risk of disease given exposure is equal to $a/(a + b)$. The odds are defined as risk/(1 – risk). Hence, the odds that an exposed person develops disease are

$$\frac{a/(a+b)}{b/(a+b)} = a/b$$

The odds that a nonexposed person develops disease are

$$\frac{c/(c+d)}{d/(c+d)} = c/d$$

Hence, the odds ratio (odds of disease in exposed divided by odds of disease in nonexposed) is

$$\frac{a/(a+b)}{b/(a+b)} \sqrt{} \frac{c/(c+d)}{d/(c+d)} = \frac{a/(a+b)}{b/(a+b)} \times \frac{d/(c+d)}{c/(c+d)} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}$$

Using the counts in Table 3-1 for sample size 1 yields the following:

1. Odds of disease given exposure = $a/b$ = 20/10 = 2.0
2. Odds of disease in the nonexposed = $c/d$ = 80/90 = 0.89
3. Odds ratio = 2.0/0.89 = 2.25 or more directly $ad/cb$ = (20 × 90)/(80 × 10) = 2.25

Using the smoking example, the following is found:

1. For children with a parent who smokes, the odds of developing colds or respiratory infection are two times the odds of not developing colds or respiratory infection. The fact that the odds are greater than 1 means that the event (colds or respiratory infection) is more likely than the nonevent (no colds or respiratory infections).
2. For children with a parent who does not smoke, the odds of developing colds or respiratory infection are 0.89 times the odds of not developing colds or respiratory infection. The fact that the odds are less than 1 means that the event (colds or respiratory infection) is less likely than the nonevent (no colds or respiratory infections).
3. For children with a smoking parent, the odds of the event (colds or respiratory infection) are 2.25 times greater than the odds for children whose parent does not smoke. The odds ratio is greater than 1.0, which means that the event is more likely in the children with a smoking parent.

Several points should be made. First, the choice of making colds or respiratory infections the event of interest was arbitrary; that is, the analysis could have proceeded with the event of interest being no colds or respiratory infections. In that case the odds are $b/a$ and the other calculations are arrived at similarly. Typically, one focuses on disease. Second, the counts we used are estimates and hence subject to sampling

error. It is possible to construct confidence interval estimates for any of the quantities as needed. Third, the odds ratio (as well as the odds) is much harder for most practitioners to understand than relative risks (and risk). For example, the risk of disease in the exposed group is 0.67 compared with 0.47 in the nonexposed; furthermore, the relative risk is 0.67/0.47 = 1.42. The latter, which leads to the statement that colds or respiratory infections are 1.42 times more likely when a parent smokes, has an immediate interpretation: In a population similar to the one from which the sample was drawn, the risk of the event is 1.42 times greater. The odds ratio (2.25) has no such immediate interpretation.

### Chart Reviews and Case Review Studies

A chart review is a careful compilation of information from a collection of cases, typically cases concerning a single disease, procedure, or other defining characteristic. A case review is just an analysis of a single case. This type of study can be useful in identifying potential relationships (e.g., in a review of 30 prenatal intensive care unit cases of $H_1N_1$ flu infection, length-of-stay data might be analyzed to determine its relationship to time to treatment). These studies are retrospective in nature and may offer insights into associations and certainly might offer clues regarding fruitful areas for additional research. One shortcoming of chart reviews is that chart data are often incomplete and, compared with a carefully designed prospective study, information on key variables may be missing. Compare this with a case-control study and you will see that the only comparisons one can make are internal to the collection of cases.

## EXPERIMENTAL DESIGNS

### Features of True Experimental Designs

True experimental designs have each of the following features: (1) manipulation or an intervention, (2) randomization, and (3) experimental control. There is also some measure or collection of outcome measures that are connected with the purposes of the study. These are defined in the study protocol and should be logically connected to the research question that motivates the study. In what follows the studies are for the most part pictured (described) as consisting of two treatment groups. This is merely for convenience (simplicity) because there is no limit to the number of treatment groups in an experimental study. Labeling the groups as "treatment group" and "control group" is also for convenience. To have randomization between groups one must have at least two groups, and typically one is the control group (a comparison

group that comprises the standard treatment); however, this may not be the case in a particular study.

The language of study design arises from several sources, two important ones being industrial/agricultural experimentation (scientific areas in which true experimental studies routinely take place) and psychology/education. The latter probably accounts for use of terms such as "pretest/posttest" designs, which are used even when the outcome measures are not strictly speaking something that could be described as a test. The former are the source for terms such as "split plot" designs, which suggest agricultural experimentation.

Previously, we discussed nonexperimental designs. These were observational in nature. Nonpejoratively they can be described as passive—the approach is basically let's see what happens. Many useful scientific conclusions are arrived at in this manner (think astronomy).

### Manipulation or Intervention

In experimental studies something is done that is intended to affect subjects in the study. It might be the administration of a drug or a change in protocol by the pharmacy or almost anything thought to affect an outcome. This intervention divides the subjects in the study into groups (defined by the intervention). A simple division is one in which one group receives a treatment (treatment group) and another group receives no treatment (control group).

### Randomization

Randomization refers to the introduction of chance into the selection or assignment of subjects to treatments. Randomization can occur at two levels: one described as random selection and the other described as random assignment. *Random selection* refers to how one draws the sample of people (subjects) for the study from a population. In most nursing and medical studies, this is desirable but not achievable; however, some studies do have random selection. The form of random selection can be simple (characterized loosely as "each subject in the population has an equal chance of selection") or more complicated (such as using stratification and clustering). These more complex methods of selection are often used in surveys. *Random assignment* refers to how one assigns the subjects in the sample that is drawn to different groups or treatments in the study. Random assignment is sufficient for a study to be referred to as experimental. Patients who enter a clinic and are diagnosed with disease A are randomly assigned to "old treatment" and "new treatment." This is not a random sample of patients from which assignment is made.

This is typically referred to as a convenience sample. However, the random assignment has made it possible to argue that different outcomes are attributable to the treatment and hence the design is an experimental one.

The purpose behind randomization is to remove the effects of systematic variation in the subjects (e.g., the sickest are assigned to treatment) that might confound (confounding factors are factors other than the intervention that might be used to explain the observed outcomes) the results. Underlying randomization theory is the notion of potential differential effects; that is, attributes of the subjects (known or unknown to the researcher) might mediate the impact of the intervention. The idea behind randomization is that these differential effects are "averaged out"; that is, all those characteristics of the subjects that impact on the outcome (measure) are distributed evening among the treatment groups. This ideal is often not achieved and in small samples is possibly unachievable. Also, the researcher must contend with "bad luck." If you are living on luck (randomization), then one result is bad luck. In practice this means that the groups (those defined by the intervention) are not similar with respect to the subjects assigned to the groups. Note that the implied "comparison" can be made explicitly only with known aspects of the treatment groups, but faith in the randomization leads the researcher to believe that those factors unknown to the researcher (or factors with respect to which no data have been collected) have been averaged out.

### Experimental Control

Control (experimental control) is concerned with consistent administration of the intervention. In simplest terms, a well-controlled experiment is one in which the treatment groups are managed identically with respect to both the administration of the intervention and any aspects of the study (factors) that might influence the outcome, both known and unknown to the researcher. Key among these is "attention control"; that is, ensuring that those in different treatment groups receive the same level of attention (including surveillance). This is frequently achieved in clinical settings by adhering to a well-designed protocol. Inclusion of blinding is a form of control. Blinding the subjects means that the subjects do not know what form of treatment they are receiving. Those administering the study may also be blinded (as well as those performing statistical analyses). Blinding is often easy to achieve in drug studies, but in other studies it is impossible (e.g., surgery versus radiation in prostate cancer).

## Notes on the Narrow Definition of Experimental Study

It is worth noting that the definition of an experimental design we are using is quite narrow and excludes certain things that one might think of as experiments. Suppose

one randomly selects depressed subjects infected with human immunodeficiency virus from an urban clinic database to study their response to a nurse-based program of laughter therapy. At the end of the study, subjects are administered an instrument that measures depression. One might learn useful information from this study, but there has been no random assignment because there is only one treatment group. This design could be characterized as a one-group-only posttest design, where posttest is shorthand for the idea that measurements are made only after the intervention. This study could be designed to allow for preintervention measurement of depression. This leads to another form of division in which subjects are measured initially (baseline), then the intervention occurs, and then subsequent to the intervention the subjects are measured again (these types of studies in the simplest form are referred to as pre–post studies; when more than one postintervention observation is made, these may be called repeated-measures studies or perhaps longitudinal studies). A one-group pre–post design is sometimes called preexperimental (and would not conform to our definition of experimental).

Laboratory or certain industrial experiments are frequently designed as one-group posttest studies when the phenomenon is well understood. For example, the breaking strength of metal wire of a certain composition might be known (actually the distribution of breaking strengths) and the researchers wish to study the effect of a change in composition (e.g., tungsten is added). Ten pieces of wire with the new composition are tested, and the breaking strength is measured with the intention of comparing these measurements with a known outcome (e.g., average braking strength). Those performing this study would consider the study to be experimental. Research on human subjects is usually more complicated than this. Furthermore, the effects of a human subject's intervention in general tend to be more varied in the response often because of idiosyncratic attributes of the subjects (that are unknown or not understood by the researcher).

## Purposes Behind the Use of Experimental Designs

Why are studies undertaken? Fundamentally, the researcher wishes to gain and communicate knowledge about some process or phenomenon. In epidemiological studies (which include clinical research and related activities) the subject is disease. With nonexperimental designs one can establish association, but it is difficult to develop compelling arguments regarding causality, although occasionally this has been done with some success. In experimental studies one can come closer to establishing cause and effect.

Cause and effect in human subject studies is not only hard to establish; in most cases it has to be carefully circumscribed because the effects of an intervention are

typically differential—that is, the intervention does not affect all subjects to the same degree or at all. For example, under controlled circumstances a Newtonian apple always falls toward earth when it falls from a tree. Every apple falls. Even in cases of perfect experimental control in a cancer study, perhaps in 50% of subjects tumor size is reduced. Not only did only some of the subjects respond to the treatment, but some responded without the treatment (because of placebo or other unknown causes). Often in intervention studies the response rate is described in terms such as these. This is not classical causality (if *x* then *y*; if not *x* then not *y*). Human subjects are extremely idiosyncratic. In addition, other causes can operate to complicate the explanation. Hence, cause and effect in studies on human subjects is more problematic. Finally, does one ever observe cause and effect in studies on human subjects? One can argue that only correlation (association) is ever observed (where correlation is defined more broadly than a linear relationship).

### Internal Validity

In assessing internal validity, one answers the question of whether the study measures what it set out to measure. Often, and more importantly, did *x* really change *y*? This issue is not relevant to most observational studies (e.g., cohort or case-control studies), although such studies have been used to argue causality (and in this circumstance the issue of internal validity arises). For intervention studies internal validity is crucial, so true experimental and quasi-experimental designs can be evaluated with respect to the degree of interval validity particular designs possess. Put simply, internal validity addresses the question of whether observed changes in outcomes can be attributed to the intervention (manipulation) and not to possible other causes. In the real world this ideal can never be completely achieved. In that sense one can think of degrees of internal validity.

Internal validity applies to a specific study. That is, when one has internal validity, then one can claim that "what was done" did affect the outcome. One can examine the construct validity of a study, which refers to the connection of your study to its theoretical constructs. For example, in a study of regular nurse visits to elderly shut-ins, the visits improved scores on a quality of life survey (compared with the control subjects, who received no visit). Properly designed, this study had high internal validity. Suppose in another study there were three arms: no visit, nurse visit, and meals-on-wheels delivers a hot lunch. In this study there is no difference between the quality of life scores for the nurse visits and the meals visits. In the first study the outcome is attributed to the nurse visit, and it is true the visit led to higher quality of life scores (hence internal validity was high), but human contact seems to be the true explanation of the effect (as established in the second study). Or was it the hot lunch?

Below are some commonly identified threats to internal validity. When examining a particular study, one can examine each of these with respect to its perceived impact on internal validity.

- *History:* Did some other current event effect the change in the dependent variable? Typically, the event occurs between a first and second measurement (between pre and post).
- *Maturation:* Were changes in the dependent variable the result of normal developmental processes? This would include any change associated with the passage of time independently of treatment.
- *Statistical regression:* Possible if groups are selected on extreme scores or other extreme characteristics. This is classically referred to as "regression to the mean."
- *Selection:* Were the subjects self-selected into experimental and control groups, which could affect the dependent variable? The usual remedy is random assignment, but this is not always effective.
- *Experimental mortality:* Dropouts and loss to follow-up. How were missing data treated in the statistical analysis?
- *Testing:* Did the pretest affect the scores on the posttest?
- *Instrumentation:* Did the measurement method change during the research?
- *Design contamination:* Did the control group find out about or interact with the experimental treatment? This includes contamination via personnel administering the "intervention," such as nurses who are involved in different arms of the study communicating.
- *Selection–maturation interaction:* The selection of comparison groups and maturation interacting that may lead to confounding outcomes and erroneous interpretation that the treatment caused the effect.

### External Validity

External validity refers to generalizability. That is, to what populations do the conclusions of a study pertain? This is not relevant in the absence of internal validity. Construct validity also relates to generalizability in as far as this measures the degree to which the theoretical constructs being measured in the study are valid. Ideally, a study should have high interval validity and high external validity. Actually defining the population to which the conclusions of a study apply is nontrivial and often likely to be overstated.

A threat to external validity is an explanation of how you might be wrong in making a generalization regarding the findings from some study. Generalizations typically

involve extending the study to different people, places, or times. These errors of generalization can occur at the researcher level (that is, those who performed the study can make invalid claims) or by users of the study (you, if you are engaged in advanced nursing practice and apply findings in a way that is inconsistent with the study's generalizability). Perceived or potential threats to external validity can be the inspiration for additional research in which the original study is replicated with new people, places, or times. Below are listed some commonly identified threats to external validity:

- Unique program features.
- Effects of selection: The pool from which random assignment was made is often a large convenience sample, not random selection from some well-defined population. This makes generalizability problematic.
- Effects of setting/situation: All situational specifics (e.g., treatment conditions, time, location, lighting, noise, treatment administration, investigator, timing, scope and extent of measurement, etc.) of a study potentially limit generalizability.
- Effects of history: If historical circumstances were different at the time the study was conducted, findings no longer apply.
- Effects of testing: In general, testing will not occur and certainly not pretesting. For example, implementation of a program (intervention) to improve quality of life probably will not include testing.
- Reactive effects of experimental arrangements: Subjects know they are in study. Another aspect of this revolves around the question of whether those that participate in studies are different from those who do not or did not have an opportunity to participate.
- The Pygmalion effect, or Rosenthal effect: The phenomenon in which the greater the expectation placed on people, often children or students and employees, the better they perform. This is related to the reactive effects of experimental arrangement points.

## Important Examples of Experimental Designs

The literature of experimental design is vast and typically is presented in the context of statistical analysis. The designs selected for discussion in this chapter are typical of those that arise in health-related studies that would be encountered by those engaged in advanced nursing practice.

### Randomized Controlled Trial

In the jargon of experimental design (as a subject area of statistics), a completely randomized design is one in which experimental units are randomly assigned to treatments. In most clinical and epidemiological studies the experimental units are human subjects, and studies (designs) of this type are typically referred to as randomized controlled trials (RCTs). For a single study these are often considered the best designs (that is, internal validity is high) or the gold standard. The other designs considered in this chapter are variants of this fundamental design. The basic outline appears in **Figure 3-5**.

This description is quite general and covers many possible variations with respect to details. Some key points are as follows:

- The sampling may take many forms, including convenience sampling. This can greatly affect external validity.
- This is a pre–post design if measurements of the response variable(s) are made at baseline (e.g., quality of life instrument is administered) and postintervention.
- Data collected may be nominal, ordinal, or interval, which affects the data analysis and the type of conclusions drawn.

Consider the following simple example, which begins with the research question: Does adding a video to predischarge education by nurses increase patients' (who have had colostomy surgery) knowledge about infection risks? In outline, the study design is as follows: (1) 1 day before discharge (and after the standard nurse conducted education) all patients are given a 10-question "test" designed to measure their knowledge of risks and good practices. (2) Those in the intervention group (randomly selected) are shown a video before discharge. All patients are given the same test at discharge.

In this example the key features of the RCT are easily identified. There is an intervention, which is the video. The treatment groups are created by randomization and consist of those who received the "standard education" and those who received the standard education plus the video. There is at least one clearly defined outcome variable that in this study is derived from the test administered pre (subsequent to the standard nurse conducted education) and post (that is, subsequent to the video). The test might be scored (e.g., percent correct answers), or individual questions on the test may be considered as outcomes. Observe that this study cannot be blinded with respect to subjects because as those receiving the intervention will certainly know they saw the video. The study protocol should clearly describe how the intervention is to be administered (e.g., is there a question-and-answer period after the video) as part of experimental control.

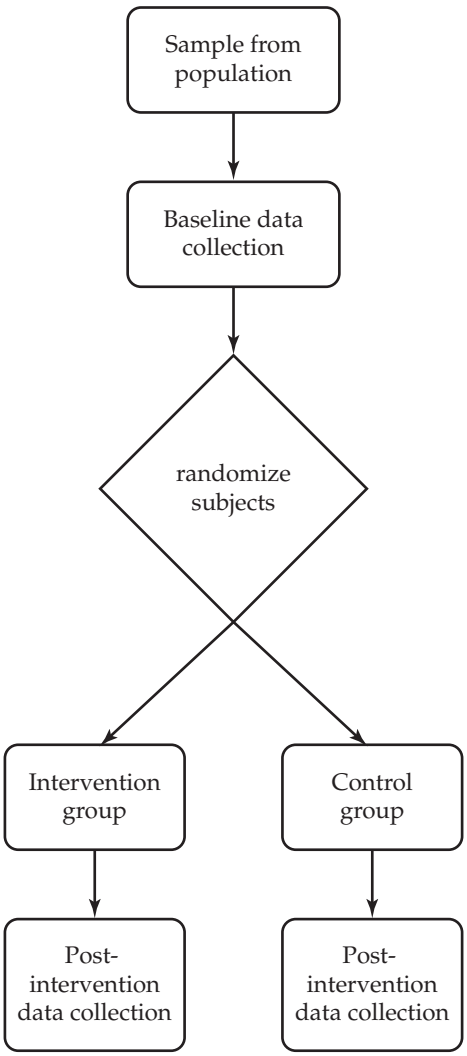


**FIGURE 3-5** Randomized controlled trial outline.

### After-Only (Posttest Only) Design

In an after-only design the subjects are randomized into treatment groups without preintervention measurement of response (outcome) measures (**Figure 3-6**). This does not mean certain data cannot be collected before randomization (these may be covariates such as age, gender, education); however, no measure of the response (outcome variable) is made preintervention. In some studies only postmeasurements may

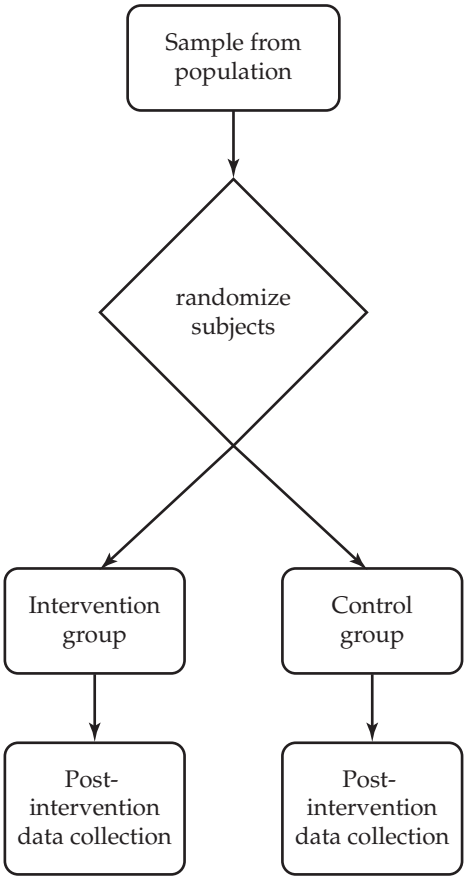


**FIGURE 3-6** Post-test design.

be taken; for example, a study of post surgical hospital readmits for groups randomized between two predischarge instruction methods for wound care. The after-only design results in data from two independent groups (a consequence of random assignment). Variability within the groups could be quite large, which means that successfully detecting differences between the groups may require large samples. In studies involving quality of life, the acquisition of knowledge, and assay or other physiological measurements, collecting preintervention measurements can increase the power associated with statistical analyses because inherent human subject variability can be reduced taking preintervention measures into account when performing analyses. However, randomization alone is a sufficient basis for declaring differences between treatment groups are associated with treatment (intervention).

### Solomon Four-Group Design

The Solomon four-group design is appropriate when the researcher suspects that preintervention measurement has an effect on the response. That is, this design considers the possibility that baseline measurement affects postmeasurements; for example, taking pretest influences posttest (typically either through learning or increased awareness). The four groups are as follows (**Figure 3-7**):

- A: Baseline data collected and subjects receive intervention (treatment)
- B: Baseline data collected and subjects are in the control group (standard treatment)
- C: No baseline data collected and subjects receive intervention (treatment)
- D: No baseline data collected and subjects are in the control group (standard treatment)

This design can be modified so that certain baseline data are collected from all four groups and other baseline data are not. That is, outcome measures data are omitted in "no baseline data" but certain other (covariate) data (e.g., age, gender, race and such) might be collected.

What is the idea behind this design? Suppose a study is intended to measure the effects of a nurse-based intervention for transitioning adolescent diabetics into adulthood. Observations are to be taken at baseline and end of study 12 months later. One measure in the study is adherence to good practices as measured by a survey instrument. Does just taking the survey (which serves to remind the subjects of certain good practices) affect behavior? One way to answer that question (and simultaneously guard against this effect) is the four-group design. By comparing groups A and C with respect to end-of-study responses with the survey, one can assess the impact of taking the survey preintervention. What is learned from comparing groups A and D at end of study? The cost of this design is greater than a classical RCT because there are four groups. One additional issue is the lack of a generally agreed-on method of statistical analysis for this design. That is, how does one model it? In itself this is not that serious a drawback because using several approaches to the resulting data can lead to answers to key research questions. There are several approaches to analyzing RCT data, but that has not reduced interest in using this design.

### Crossover

In the most elementary version of a crossover design, all subjects in the study receive all treatments. This design closely corresponds to the counterfactual ideal in epidemiological studies in which, instead of comparing disease outcomes for those who are

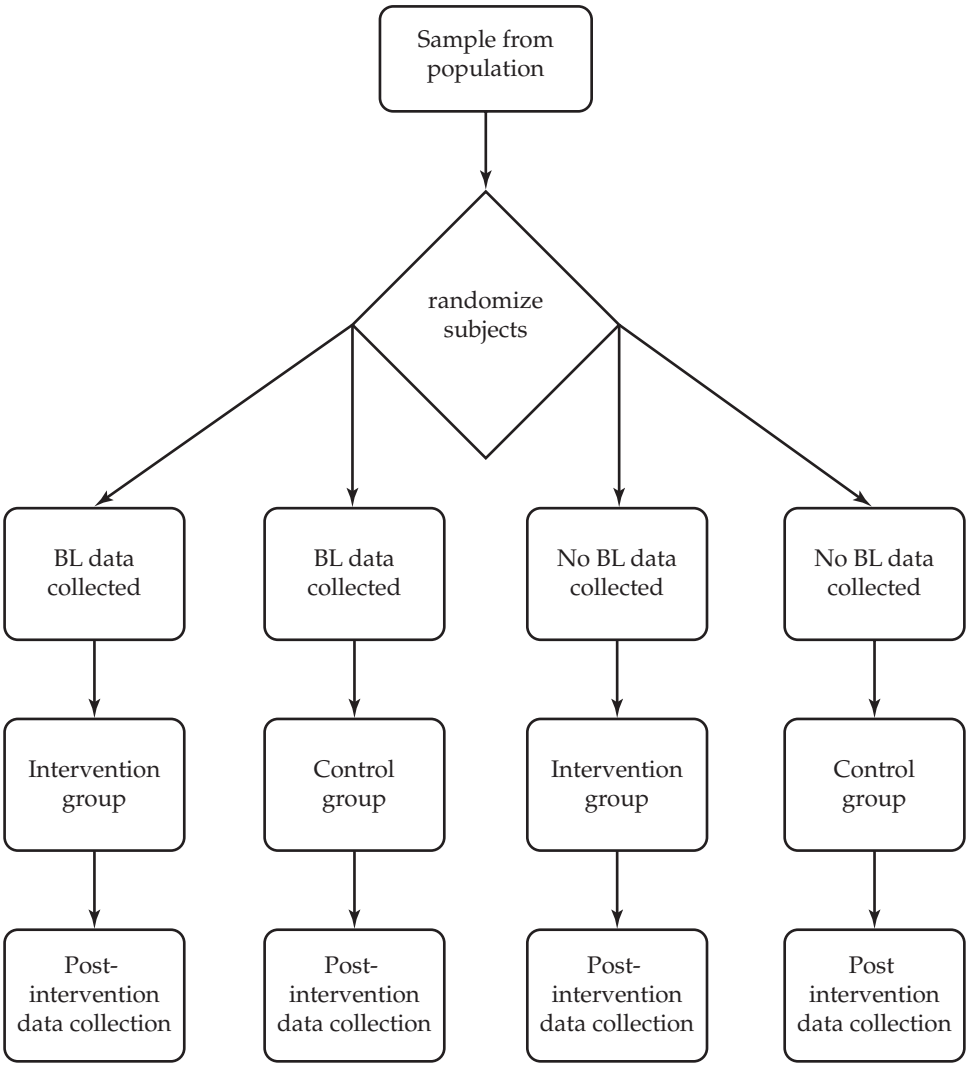


**FIGURE 3-7** Solomon four-group design. *Note:* BL, baseline.

exposed and those who are not exposed, the same individual experiences both exposure and nonexposure. Crossover designs are often used in clinical trials. The study is divided into periods, and patients receive different treatments in these periods. There can be several periods, but in the simplest case there are two. Compare this with the previous parallel designs in which the (two) treatment groups simultaneously receive

treatment (this is an abstraction because in practice patients are typically enrolled over time). In the crossover design there is not a separate comparison group. Each subject acts as his or her own control.

Because each subject acts as his or her own control, there is no possibility of covariate imbalance. However, in more complicated versions this can arise (i.e., not all subjects receive all treatments). Randomization enters the picture with respect to the order in which the treatments are administered. Types of randomization can arise in other circumstances. The rule of thumb is when in doubt, randomize.

Consider a study in which two arthritis drugs are being compared (one might be a placebo). Then each subject can be described by his or her treatment sequence: AB (drug A first and drug B second) or BA. Patients are randomly assigned to a treatment sequence. A concern in such studies is carryover effects. These are dealt with in part by having a "washout" period between treatments. For each subject (1) a drug, based on the subject's treatment sequence, is administered for some period of time and outcome measures are recorded (this might be something as simple as the subject reporting the treatment as successful [S] in controlling symptoms or unsuccessful [U]); (2) a washout period intervenes; and (3) the second drug in the treatment sequence is administered and outcome measurements are recorded.

The analysis typically considers the effect of order and a comparison of outcomes between the two treatments (and carryover effects). In the example above there are two sequences (AB and BA) and associated with each are four outcomes: SS, SU, US, and UU (using only this simple dichotomous measure, in which SS refers to success in period 1 followed by success in period 2). A response profile contains the counts for each outcome (**Table 3-4**). The goal here is not analysis of these data. However, the data in Table 3-4 do suggest certain conclusions.

From the example it is easy to see that this design can be extended to more than two treatments (which results in longer treatment sequences), that the subjects can be stratified before assignment (e.g., in this example by gender), or that for more than

**Table 3-4 Crossover Data**

| | Response Profile | | | | |
|---|---|---|---|---|---|
| **Sequence** | **SS** | **SU** | **US** | **UU** | **Total** |
| AB | 4 | 15 | 5 | 6 | 30 |
| BA | 5 | 4 | 16 | 5 | 30 |

*Note:* S, successful; U, unsuccessful.

two treatments not all subjects are assigned to each treatment (e.g., suppose there is a placebo in the previously mentioned drug study, then one might assign subjects to any of six treatment sequences such as AP, BP, AB, and so on).

### Factorial Designs

Below are the key attributes of a factorial design:

- Factors are variables that affect the outcome/response that one can set (control) during the experiment (study). One might think of these as design variables.
- Factors occur at levels (that is, they can be set at different values, which may be numerical or categorical). That is, factor A may be at $i$ levels, factor B at $j$ levels, and factor C at $k$ levels. In this case there are $i \times j \times k$ factor level combinations. Setting (or deciding on) the factor levels is the manipulation part of the design—that is, they define the intervention.
- In a factorial study the effects of the factor levels on the response are investigated by setting the levels of all factors and then observing the response. In a *full factorial* responses are observed for each factor level combination. In a balanced design with $n$ replicates there are then $i \times j \times k \times n$ observations in the sample. The number might be large, which is one of the drawbacks of full factorial designs.
- In factorial designs one examines what are called main effects and interactions. The fact that interactions can be studied explains why factorials are superior to "one factor at a time" studies.

The last bullet mentions interaction. Interaction occurs when the effect of factor A on the response is influenced by the level of factor B. Interaction is often interesting and always (when present) complicates the description of the relationship between the factors and the response.

When one adopts a factorial design, there is some number ($M$) of distinct factor level combinations. In the case of three factors described above, there are $M = i \times j \times k$ factor level combinations and the $N = i \times j \times k \times n$ subjects are randomly assigned to factor level combinations (random assignment is essential for this to be an experimental design). For example, suppose that diabetes control is the focus of a nurse-administered diet/exercise study that involves three diets and four exercise programs. There are 12 combinations of diet/exercise. If 60 subjects are enrolled in the study, then 5 subjects can be randomly assigned to each treatment combination (leading to a balanced $3 \times 4$ factorial design). Suppose the outcome measure is $HbA_{1c}$. If the subjects are measured at baseline and at end of study (e.g., 6 months), then this is really

just a somewhat fancy RCT. It is worth noting that if the outcome measure were a dichotomous variable with values "in control" and "not in control," then it is unlikely that a study with 60 subjects would have enough power to make the study worth conducting.

The benefit derived from the factorial design is that interaction between diet and exercise can be examined (and quantified). When interaction is absent, then the main effects of diet and exercise program can be evaluated independently of one another.

The example above is a full factorial. There are designs called fractional factorials from which one can gain important information with less experimental effort (*N* less than 60). Authors such as Montgomery (2009) provide a thorough discussion of these approaches. One type of factorial design that could be a great value in clinical studies (it is quite commonly used in industrial applications) is screening designs. In screening designs several factors (sometimes as many as seven or eight) are set at only two levels (such as high or low or yes or no), leading to what are called $2^k$ designs (there are $k$ factors). The goal is to identify "active" factors (that is, ones that influence the outcome).

In principle, factorial designs are separate from post-hoc stratification using several factors in as far as these factors (factors other than the design factors) are not subject to random assignment and are not "design variables," the levels of which are set by the researcher with the design. Analyses of this type are often conducted using analysis of variance (ANOVA). For example, in the previous study one might simultaneously look at the effects of gender and race on the outcome using two-factor (unbalanced) ANOVA (actually a general linear model)—the exact role of the two design factors (diet and exercise) in the model might complicate the analysis but does not obscure the point. In this case association of gender and race with $HbA_{1c}$ can be established, but a causal argument such as the diet/exercise one is more tenuous.

## QUASI-EXPERIMENTAL DESIGNS

As the name suggests, quasi-experimental designs share some of the characteristics of true experimental designs. The purposes (primarily to investigate cause-and-effect relationships) are similar to those motivating true experimental designs. One or more characteristics of true experimental designs are missing, either because they cannot be achieved or because there are valid reasons why a quasi-experimental design might be superior.

### Nonequivalent Control Group Design

The fundamental difference between this design and the classic RCT is that there is no random assignment to treatment groups. There are still comparison groups and

experimental control still needs to be exercised (apart from randomization). Most frequently, the treatment groups arise naturally, for example, they may be different hospitals, different nursing facilities, different practices, and so on. **Figure 3-8** contains a schematic for this design. It is important to note the following:

- Sampling is going on, that is, the subjects in the two groups came from somewhere. This is true for experimental designs also. This impacts on external validity—generalizability—regardless of whether random assignment takes place.
- The most convenient "assumption" is that the two groups are similar/comparable at baseline. This should be investigated (confirmed) using statistical methods. Sadly, this implies that only known confounding factors can be examined (or, worse yet, using only factors on which data can be collected). For comparison, consider that in designs in which random assignment is used it is often assumed that baseline differences in key (active) variables (as well as unknown variables) have been averaged out. In practice this may not occur,

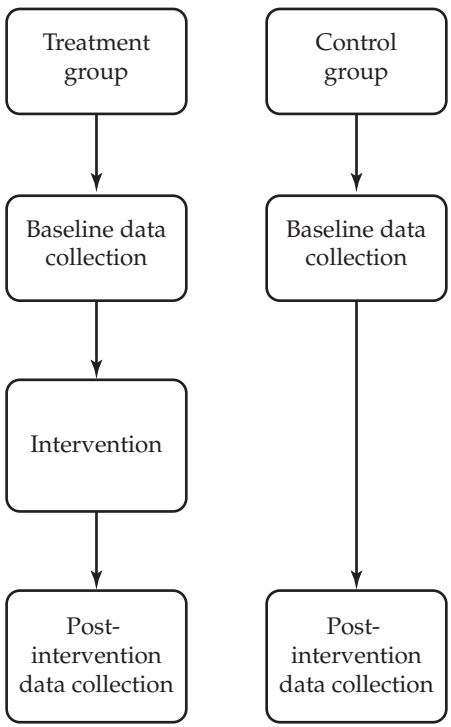


**FIGURE 3-8** Nonequivalent control group design.

and in the context of the number of studies actually conducted, many will fail to be properly randomized.

In circumstances in which the groups are different, statistical methods (such as analysis of covariance) may be used to control for this. The remarks in the preceding bullet regarding known versus unknown factors apply here also. It is with regard to unknown or unanticipated factors that nonequivalent control group designs may lead to incorrect conclusions.

This design can be appropriately used when contamination is a danger. In a nursing setting, for example, the intervention can be used in one ward/hospital/unit and the comparison treatment in another so that nurses and patients in the two arms do not interact.

Consider the following example based on the research question: Does an additional "hands-on" session regarding baby care with a nurse while music is played improve a new mother's belief that she is ready to take care of her baby? A (hypothetical) validated instrument (BC10) consisting of 10 Likert (ordinal) scale questions is used to measure the mother's belief in her readiness. Note that the nurses will have to be trained and prepared for the hands-on session. To avoid "contamination" the researchers decided not to randomize patients within a unit, so two units were used for the study: One unit was randomly chosen to be the intervention unit. The nurses on that unit were trained to perform the intervention. The instrument (BC10) was administered after the "standard" baby care instructions were given to the mothers (usually day 2 of their hospital stay). In the intervention group the mothers were given the hands-on session. Before discharge the instrument was readministered to all subjects.

If the researcher wants to separate the "hands on" and music effects, four groups (units) are needed. Defined how? If one is concerned about a test–retest effect, then one could modify the Solomon design (without randomization) to suit this situation.

Selection bias is a major concern with this design. One method for detecting this is to have multiple (two or more) preintervention tests (measurements) to allow the analyst to detect trends in the measurements (and assess maturation threats). When there are differential preintervention maturation effects (different trends), that is suggestive of selection bias.

## After-Only Nonequivalent Control Group Design

For the after-only nonequivalent control group design (**Figure 3-9**), no baseline data are collected (on the outcome variable). Because there are no baseline measurements, only the treatment groups can be compared postintervention, which makes it analo-

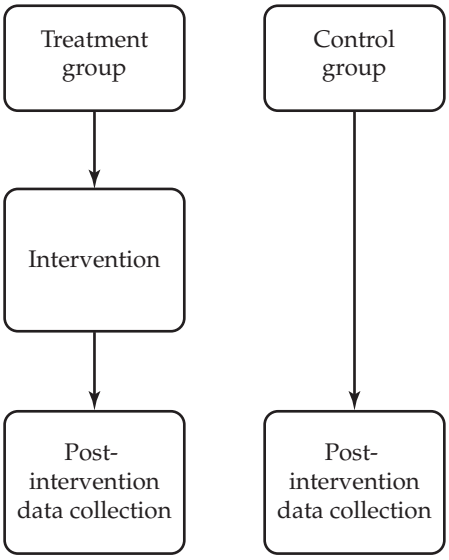


**FIGURE 3-9** After-only nonequivalent control group design.

gous to its after-only experimental counterpart. The groups must be similar, or one must control for these differences to make valid comparisons and conclusion regarding the effect of the intervention/treatment. Consider an example based on the following research question: Does a nurse manager in a medical home plan produce better outcomes in adolescent diabetes patients? Two medical practices agree to participate in the study, and one is randomly chosen to have a nurse manager. The treatment group has 22 patients and the control group has 27 patients. After 6 months $HbA_{1c}$ levels are measured for the 45 patients available for follow-up. Would this design be improved with pretest (baseline) measurements? In what way? Validity? Power?

## Single-Group Designs

Single-group designs do not include a comparison group. For designs in which measurements on the outcome variable are made once preintervention and once postintervention, each subject is his or her own control and the design is frequently described as a single-group pretest posttest design. This is just a two-time point repeated measures design. If there are multiple postintervention time points at which outcome measurements are made, then this is single-group repeated-measures design. The key

point is that repeated measurements on the same subject are likely to be nonindependent, which complicates the analysis. In this type of study interest is in the evolution of the outcome measures over time.

In a design in which there is a single group with after-only measurement (that is, there are no preintervention measurements), the researcher can make comparisons with respect to the intervention that are external to the study. For example, 5-year survival rates (actually proportions) for a treatment can be compared with published survival rates. Ascribing the observed survival rates to the treatment is tenuous because the outcomes may be solely attributable to characteristics of the treatment group or, worse yet, the result of some exogenous event. For example, in a study on sexual behavior in teens, behavior is measured baseline (how?), a nurse-based educational intervention is administered, and subsequently behavior is measured again. Between the pre- and postmeasurements there was a television series on the danger of sexually transmitted infections. Exogenous events such as this can contaminate the study. Minimally, critics can argue that unknown events induced the outcome (e.g., there was a full moon). Single-group after-only studies are typically exploratory and are useful for estimating characteristics of a population of interest but are unlikely to produce convincing cause-and-effect arguments unless supported by follow-up studies with a comparison group.

## REPEATED-MEASURES (TIME SERIES) DESIGNS

Time series designs can be repeated-measures designs when the experimental unit is a human subject. In modeling some epidemiological measurement over time (e.g., infant mortality), what are frequently referred to as time series methods might be used. Modeling of this type is not the subject here.

Repeated-measures designs are sometimes referred to as longitudinal studies. The essential idea is that measurements are taken over time, perhaps at regular intervals of time or perhaps not. Because the repeated-measures design is so frequently used, it is useful to summarize some of its features even though some of this repeats previous discussions in this chapter.

- Repeated-measures designs may be experimental, quasi-experimental, or nonexperimental.
- Repeated-measures designs may have an intervention or not. If there is an intervention in which only some subjects receive the intervention, then the intervention groups define a between-subjects effect. There may be more than two treatment groups.

- Often, there is interest in a time effect. That is, the measurements evolve over time (even when there is no intervention and hence a single group). This is sometimes referred to as "maturation" but is more often referred to as a within-subjects effect (a statistician's characterization).
- When there is more than one arm/group (e.g., a treatment and a control arm), the interaction between time and group is often the focus of the study. That is, is the evolution of the repeated measurements of the outcome variable the same for the treatment group and the control group? This idea extends to more than one treatment group.
- The pretest–posttest design is a simple example of repeated measures. When there are only baseline and end of study measurements, then change = baseline – end of study is often used to measure the time effect.

## STUDIES OF STUDIES: LEVELS OF EVIDENCE

In advanced nursing practice one is often interested in "degrees of proof" or levels of evidence. Once a design has been chosen (or even when not consciously chosen) and a study completed, the study will result in claims regarding the nature of some process or phenomenon. In our case some disease- or exposure-related conclusion will be offered. The truth offered will usually be supported by statistical and other arguments. What level of belief should be attached to these conclusions? And if there is belief, to whom (that is, to what population of subjects) do they apply? The practitioner asks, "can I use this knowledge in my practice?" Assessment of internal validity of a study offers a partial answer. That focuses on design. Generally, RCTs are thought to have high internal validity. Next, one might examine effect sizes. That is, given statistical significance, what are expected effects associated with a particular intervention? Recall that statistical significance is an artifact of three things: true effect size, sample size, and luck. The latter appears in two contexts: (1) a small (or nonexistent) true effect in the population produced through randomization a large observed effect in the treatment group and statistical significance and (2) a large effect through randomization leading to roughly equivalent treatment groups and hence lack of significance. Examination of a single study, no matter how carefully designed, never allows one to know whether luck has intervened in a manner that leads to false conclusions. Observe that published studies with false conclusions tend to be those of type 1. If 5% level of significance is used, then inevitably type I errors will arise and the practitioner will never know whether the study under consideration is a type 1.

Apart from the preceding concerns, having belief in good luck and observing a clinically significant effect size (perhaps by examining a confidence or tolerance

interval), the practitioner must still ask to whom does this apply? That is, does this apply to my patients? Questions of external validity are even thornier than those of internal validity, and one must rely on the description of the study population (typically a sample with certain characteristics) to address this. Very seldom are studies based on random samples from some well-defined population. More typically, they are some sort of convenience sample.

One approach to the problems raised here is to engage in a systematic review. That is, one critically examines papers relating to the research question. Finding published papers if they exist is usually relatively easy. But what about studies that were not published? Of special interest are those not published because the findings were not statistically significant. And what is their role in an overall assessment of the "truth" regarding some intervention (or more generally a description of some disease phenomenon)? In general, there is belief that a systematic review (if properly conducted) offers a higher level of evidence regarding the truth than a single study. There is no doubt that it is more nuanced. When statistical methods (as opposed to clerical or ethnographic methods) are used to "combine results," the result is called meta-analysis. Meta-analytic studies are often considered to offer the highest level of evidence regarding the truthfulness of claims and are sometimes referred to as a "state-of-science" assessment. Meta-analytic methods are quite popular, but combining results from different studies is a nontrivial undertaking; given the same set of studies, different researchers might reach somewhat different conclusions regarding "the truth." Therefore, some caution should be exercised in interpreting meta-analytic results.

## KEY POINTS

- Study design can be viewed from two perspectives. First, from the point of view of the primary researcher, careful consideration of study design is needed to ensure that methods and resources are combined in a manner that facilitates answering the research question motivating the study—that is, will the study provide answers to the research question in a manner that will convince scientists and practitioners that the conclusions of the study are useful and valid? Second, from the point of view of the practitioner (e.g., someone engaged in advanced nursing practice) who might implement findings from a study, critical examination of the research design is essential in evaluating the findings with respect to effects and validity. Put more bluntly, can these research findings be put to use in serving the needs of the people to whom I wish to apply them?

- Although there is a general consensus that the strongest causal evidence arises from true experimental designs (the gold standard being RCTs), the designs discussed in this chapter have meaningful roles and are often the best or only approach available either because of dollar/resource constraints or practical or ethical considerations.
- Cohort studies are the basis for classical epidemiological studies. They are expensive, and prospective ones take a very long time to conduct. Being observational studies, it is quite difficult to develop convincing cause-and-effect arguments based on them. However, cohort studies do offer clear opportunities to order events in time so that relationships between exposure and disease can be discovered and elucidated upon in ways experimental designs cannot offer.
- With respect to their frequency of use, case-control studies have turned out to be quite useful despite limitations regarding what can be estimated (inferred) with this design. These designs are the easiest way to study disease that either develops slowly or is relatively rare.
- Intervention studies, both experimental and quasi-experimental, have been invaluable sources of scientific information and knowledge. The essential methodology used in well-designed intervention studies is comparative and strives to answer the question, did the intervention work when compared with some other treatment? Comparison of this type is a fundamental part of the scientific method. Design principles, along with appropriate statistical analysis and modeling, are what make comparative studies effective in discovering truth. Aspects of design such as (random) selection, (random) assignment, blinding, experimental control through a carefully worked out protocol, pretesting, and fidelity are the major influences on validity.

## CRITICAL QUESTIONS

1. Odds ratios can be appropriately calculated for data from exposed cohort, case-control, and cross-sectional studies. In which of these can one calculate (estimate) relative risk? Which of these measures of risk is easier to explain to a patient (odds ratio versus relative risk)? In which of these can one calculate attributable risk? Which of these measures of risk is easier to explain to a patient (relative risk, attributable risk, or the odds ratio)?
2. Create three scenarios for studies in which it is not feasible or prudent to collect "pretest" data.

3. Why is it okay to compare after-only outcomes in an experimental study but not in a quasi-experimental study?
4. Conduct a mind experiment in which your research question cannot be answered using cohort study data. Repeat this for an experimental study.

## REFERENCES

Garson, G. D. (updated 2010). Research design. In Statnotes: Topics in multivariate analysis. Retrieved from http://faculty.chass.ncsu.edu/garson/PA765/design.htm

Montgomery, D. C. (2009). *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons.

Rothman, K. J. (2002). *Epidemiology: An introduction*. New York: Oxford University Press.

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

Shindler, E. (updated 2010). Framingham Heart Study. Retrieved from http://www.framinghamheartstudy.org/index.html

Trochim, W. M. (updated 2006). Research methods knowledge base. Retrieved from http://www.socialresearchmethods.net/kb/design.php