Epidemiology Primer

By Gregory B. Tallman, Pharm.D., MS, BCPS, BCIDP

Reviewed by Micheline A. Goldwire, Pharm.D., MA, MS, BCPS; Sarah J. Billups, Pharm.D., BCPS; and Kevin Reynolds, Pharm.D., BCPS

LEARNING OBJECTIVES

1. Evaluate epidemiologic measures of frequency and measures of association.
2. Distinguish elements of epidemiologic studies intended to reduce bias or address potential confounding.
3. Analyze measures of association and outputs of analytical models used to control for confounding.
4. Assess the adequacy of study methods to address potential sources of bias and confounding.
5. Justify the use of causal inference frameworks in the interpretation of study results.

ABBREVIATIONS IN THIS CHAPTER

IPTW	Inverse probability of treatment weighting
RCT	Randomized controlled trial

Table of other common abbreviations.

INTRODUCTION

Epidemiology is "the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control of health problems" (Porta 2008). Widely considered one of the core sciences of public health, epidemiology is nevertheless a relatively young field of study that has largely developed over the past century. In that short time, many concepts in epidemiology have made their way into evidence-based medicine as principles that form the foundation of how pharmacists evaluate the medical literature and assess the validity of medical studies.

In recent years, new developments in epidemiology have clarified concepts related to bias and causal inference, but these ideas have not disseminated into applied evidence-based medicine as much as newer statistical tools to control for confounding, which creates the potential to misuse these statistical methods. Pharmacists must be knowledgeable of current concepts and methods in epidemiology in order to stay up to date with the current medical literature and provide the best patient care. Given how epidemiologic principles influence evidence-based medicine, some of the terms and concepts in this chapter may be recognizable to pharmacists. However, sometimes, the same concept can go by many names, and the same term can mean different things to epidemiologists. In this chapter, we review fundamental and emerging concepts in epidemiology before looking more closely at the tools an epidemiologist keeps in the toolbox and some of the common pitfalls in applying these tools to data.

CONTEMPORARY CONCEPTS IN EPIDEMIOLOGY: THINKING LIKE AN EPIDEMIOLOGIST

Common Epidemiologic Measurements

At its most basic level, epidemiology is grounded in the act of counting within groups, and the fundamental epidemiologic measures found in epidemiologic research, though unchanged in recent years, remain foundational to the contemporary ideas and methods emerging in the field (Lash 2021). A pharmacist seeking to understand literature relying on epidemiologic methods should have a strong understanding of these measures, including what they are, when different measures are appropriate to report, and how to interpret them.

This chapter focuses on using common epidemiology measures in the context of describing health outcomes. In addition to quantifying many different health outcomes (e.g., onset of a disease or condition, death, clinical cure), clinicians may want to measure other health determinants such as exposures, interventions, or risk factors for an outcome. The measures described hereafter can be applied both to outcomes and to these other health determinants. Furthermore, we may use treatment and intervention interchangeably with exposure and disease interchangeably with outcome.

BASELINE KNOWLEDGE STATEMENTS

Readers of this chapter are presumed to be familiar with the following:

- General knowledge of common study designs and research methods and associated terminology
- General knowledge of interpretation of p values and confidence intervals and associated terminology
- General principles related to data collection, measurement, and probabilities and associated terminology

Table of common laboratory reference values

ADDITIONAL READINGS

The following free resources have additional background knowledge on this topic:

- Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951-67.
- Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.
- Alexander LK, Lopes B, Ricchetti-Masterson K, et al. ERIC Notebook.
- Vetter TR, Schober P, Mascha EJ. Biostatistics, epidemiology and study design: a practical online primer for clinicians.

Measures of Frequency

Before making any conclusions about a disease or outcome, we must first quantify the outcome of interest. Epidemiologists or clinical researchers may wish to measure how widespread a risk factor or health outcome is in the population, which can be achieved by measuring the prevalence of a disease. Or they may wish to describe the occurrence of new cases of a health outcome, which can be achieved by measuring the incidence of a disease.

Prevalence

Prevalence describes the degree to which a risk factor or health outcome is present in a target population at a given point in time and therefore reflects the existing burden of the outcome of interest (Lash 2021; Westreich 2019a). Prevalence is most useful for measuring chronic diseases (e.g., asthma) because acute conditions (e.g., infections) are less likely to be captured as prevalent cases; however, point prevalence surveys have been conducted to estimate the burden of acute conditions such as health care–associated infections (Magill 2014). When calculating prevalence, a prevalent case is an individual with the risk factor or outcome at the time when prevalence is estimated. The most common way to express prevalence is the prevalence proportion (Box 1), which can range from 0 to 1 (0%–100%). Prevalence may also be reported as odds (prevalence odds) or simply as counts

Box 1. Common Epidemiologic Measures of Frequency[a]

$$\text{Prevalence proportion} = \frac{\#\ \text{of prevalent cases}}{\text{total population}}$$

$$\text{Prevalence odds} = \frac{\#\ \text{of prevalent cases}}{\#\ \text{of non-cases in sample}}$$

$$\text{Incidence proportion (“risk”)} = \frac{\#\ \text{of incident cases}}{\text{total population}}$$

$$\text{Incidence rate} = \frac{\#\ \text{of incident cases}}{\text{total person-time at risk}}$$

$$\text{Incidence odds} = \frac{\#\ \text{of incident cases}}{\#\ \text{of non-cases in sample}}$$

[a]See the text for additional information.

Information from: Alexander LK, Lopes B, Ricchetti-Masterson K, et al. Common measures and statistics in epidemiological literature. ERIC Notebook 2014;2:1-5; Lash TL, VanderWeele TJ, Haneuse S, et al. Modern Epidemiology, 4th ed. Lippincott Williams & Wilkins, 2021; Rothman KJ. Epidemiology: An Introduction, 2nd ed. Oxford University Press, 2012; Westreich D. Epidemiology by Design: A Causal Approach to the Health Sciences. Oxford University Press, 2019.

(i.e., number of people with the disease), though these are more difficult to interpret and therefore less commonly used. Because prevalence captures existing cases out of the total population, it cannot provide information about the risk of developing a disease or the rate at which new cases will occur. For that, a different measure is needed.

Incidence

In contrast to prevalence, incidence quantifies the occurrence of new cases that arise over a specified period (Lash 2021; Westreich 2019a). An incident case is any individual who transitions from one state of the outcome to another (e.g., from uninfected to infected). Another crucial difference is that incidence should only be measured in a population of individuals who are at risk of developing the outcome of interest; individuals who already have the outcome at the start of the period (i.e., existing or prevalent cases) or who otherwise could not experience the outcome should be excluded. For example, if aiming to measure the incidence of prostate cancer, the population at risk should exclude individuals known to already have prostate cancer, as well as individuals without a prostate (cisgender women, transgender men, individuals with surgically absent prostates) who could thus not develop prostate cancer.

Incidence can further be divided into two distinct measures: incidence proportion and incidence rate. However, many terms are often used interchangeably for each measure of incidence, which can be a source of ambiguity when discussing "incidence." The first measure of incidence is the incidence proportion, which is the number of incident cases out of the total number of individuals at risk (Lash 2021; Westreich 2019a) (see Box 1). Incidence proportion, which may also be called "cumulative incidence" or "risk," can be interpreted as the probability of developing the outcome of interest over a specified period. As a proportion, the incidence proportion will always fall between 0 and 1 (0%–100%). When discussing "risk" (e.g., risk of an event occurring), most epidemiologists and clinicians are referring to the incidence proportion. Interpreting an incidence proportion requires knowledge of the time over which the risk was assessed. Consider the outcome of death: when measured over a long-enough time scale, the risk of death for humans is 1 (100%). Thus, a "2% risk of death" might be interpreted very differently if that risk is over the next 10 months versus the next 10 years. Hence, incidence proportions should always be reported with reference to the time of follow-up.

The second measure of incidence, incidence rate, incorporates the element of time directly into the measure by capturing not only whether the event occurred, but also when it occurred (Lash 2021; Westreich 2019a). Consequently, the incidence rate can be conceptualized as a measure of the frequency at which an event is occurring within the population. The incidence rate is also called the incidence density. The incidence rate is calculated as the number of incident cases out of the total time individuals were at risk, often expressed in person-years (see Box 1). For example, if five hypothetical patients were followed for 1, 1, 2, 3, and 5 years, respectively, they would contribute a total of 12 person-years to the incidence rate. If one individual developed the outcome, the incidence rate would be 1 case/12 person-years, or 0.083 cases/person-year. Often, incidence rates are reported at a scale that is easy to interpret and appropriate for the context; often, that is per 100 or per 1000 person-years (e.g., 0.083 cases/person-year × 1000 person-years = 83 cases/1000 person-years).

Unlike proportions, the incidence rate can range from 0 to infinity (Lash 2021; Westreich 2019a). Because of this, the incidence rate can accommodate any number of events and is useful for measuring events that occur more than once per individual (recurrent events). In contrast, the incidence proportion cannot exceed 1 (100%); thus, if the number of events exceeds the number of individuals, it is no longer interpretable in the usual sense. In addition, because time is in the denominator of incidence rates, it can be measured even when not all patients are followed for the same fixed time (a requirement for interpreting incidence proportions). Unequal follow-up time can result from loss to follow-up or when study populations are defined according to a temporary factor, such as geographic location.

Incidence may also be reported as incidence odds (see Box 1) or as counts of incident cases (Westreich 2019a). Incidence odds are not often described, though they are used to calculate odds ratios, which are widely reported. The choice of incidence measure depends on the question being asked, the study design, and the study population (Lash 2021; Westreich 2019a). Questions about the risk of disease occurring are addressed by the incidence proportion, whereas the incidence rate better answers questions about when events occur. Recurrent events are usually measured using incidence rates to capture the overall recurrence rate. However, incidence proportions could be calculated for each occurrence (e.g., risk of first *Clostridioides difficile* infection [CDI], risk of first CDI recurrence, risk of second CDI recurrence); such an approach could be used for questions about risk per occurrence. In addition, incidence rates are preferred to proportions when evaluating populations that change over time or those with extensive loss to follow-up.

Measures of Association and Measures of Effect

The preceding measures of prevalence and incidence are essential ways in which epidemiologists quantify the distribution of health-related states and determinants in a population. However, to understand whether some factor is a determinant of a health-related state, researchers must move beyond describing distributions of events through measures of frequency and begin exploring potential associations between exposure and outcome, including the presence, direction, and magnitude of these potential associations. Measures of association are generally based on comparisons

A	Diseased	Not diseased
Exposed	A	B
Not exposed	C	D

B	Diseased	Not diseased	Person-time
Exposed	A	B	PT_1
Not exposed	C	D	PT_0

Figure 1. Standard (A) and modified (B) contingency tables for calculating measures of association.This modified table incorporates PT for calculation of incidence rates.

PT = person-time.

Information from: Alexander LK, Lopes B, Ricchetti-Masterson K, et al. Common measures and statistics in epidemiological literature. ERIC Notebook 2014;2:1-5; Lash TL, VanderWeele TJ, Haneuse S, et al. Modern Epidemiology, 4th ed. Lippincott Williams & Wilkins, 2021; Rothman KJ. Epidemiology: An Introduction, 2nd ed. Oxford University Press, 2012; Westreich D. Epidemiology by Design: A Causal Approach to the Health Sciences. Oxford University Press, 2019.

of incidence proportions or incidence rates between exposed and unexposed groups; contingency tables can easily be used to calculate many measures of association (Figure 1). If an exposure is likely to have a true causal effect, measures of association are called measures of effect, though we use these terms interchangeably in this chapter.

Risk Difference

The risk difference is the difference in risk between exposed and unexposed groups (Box 2); here, "risk" refers to the incidence proportion (Lash 2021; Westreich 2019a). This measure is also called attributable risk, excess risk, and absolute risk reduction; all of these terms imply directionality to the relationship between exposure and outcome. The risk difference can range from -1 (-100%) to 1 (100%), with 0 (0%) representing the null value (the value when the risk is the same in both groups). Negative risk differences occur when the exposure is protective (i.e., exposure decreases risk of the outcome), and positive risk differences occur when the exposure is harmful. Because the risk difference measures the change in risk on an absolute scale and is calculated from incidence proportions, it should be reported and interpreted in a manner that makes clear the scale and time over which incidence was determined (e.g., "the absolute risk of myocardial infarction over 5 years was 5% higher in the exposed group") (Westreich 2019a). Statements without these details (e.g., "the risk of

Box 2. Common Measures of Association

Measures based on incidence proportion ("risk," *IP*):

- *Risk difference* $(RD)^a = IP_{exposed} - IP_{unexposed} = \frac{A}{A+B} - \frac{C^b}{C+D}$
- *Risk ratio* $(RR) = \frac{IP_{exposed}}{IP_{unexpsed}} = \frac{A\,/\,(A+B)}{C\,/\,(C+D)}$

Measures based on incidence rate (*IR*):

- *Rate difference* $= IR_{exposed} - IR_{unexposed} = \frac{A}{PT_1} - \frac{C}{PT_0}$
- *Rate ratio* $(IRR) = \frac{IR_{exposed}}{IR_{unexpsed}} = \frac{A\,/\,PT_1}{C\,/\,PT_0}$

Measures based on incidence odds:

- *Incidence odds ratio* $(OR) = \frac{Odds_{exposed}}{Odds_{unexposed}} = \frac{A\,/\,B}{C\,/\,D} = \frac{AD}{BC}$

Measures based on prevalence[c]:

- *Prevalence difference*[a] $= Prev_{exposed} - Prev_{unexposed} = \frac{A}{A+B} - \frac{C}{C+D}$
- *Prevalence ratio* $= \frac{Prev_{exposed}}{Prev_{unexpsed}} = \frac{A\,/\,(A+B)}{C\,/\,(C+D)}$
- *Prevalence odds ratio* $= \frac{PrevOdds_{exposed}}{PrevOdds_{unexposed}} = \frac{A\,/\,B}{C\,/\,D} = \frac{AD}{BC}$

Other measures for protective exposures:

- *Number needed to treat* $(NNT) = \frac{1}{|Risk\ difference|}$
- *Relative risk reduction* $(RRR)^a = 1 - risk\ ratio$

Other measures for harmful exposures:

- *Number needed to harm* $(NNH) = \frac{1}{|Risk\ difference|}$
- *Attributable fraction*[a] $= \frac{(risk\ ratio - 1)}{risk\ ratio}$

[a]These measures are often multiplied by 100 and reported as a percentage.

[b]*A*, *B*, *C*, *D*, PT_1, and PT_2 refer to cells from a contingency table (e.g., those in Figure 1).

[c]The calculations from a 2 × 2 contingency table are the same for measures of association based on incidence and prevalence. Therefore, the reader must know which measure the study reported.

Information from: Alexander LK, Lopes B, Ricchetti-Masterson K, et al. Common measures and statistics in epidemiological literature. ERIC Notebook 2014;2:1-5; Lash TL, VanderWeele TJ, Haneuse S, et al. Modern Epidemiology, 4th ed. Lippincott Williams & Wilkins, 2021; Rothman KJ. Epidemiology: An Introduction, 2nd ed. Oxford University Press, 2012; Westreich D. Epidemiology by Design: A Causal Approach to the Health Sciences. Oxford University Press, 2019.

myocardial infarction was 5% higher in the exposed group") are ambiguous and could be interpreted on an absolute or ratio scale.

Risk Ratio and Rate Ratio

Associations between exposure and outcome can also be measured on the ratio scale. The risk ratio is the ratio of the incidence proportion in the exposed group to the incidence proportion in the unexposed group, whereas the incidence rate ratio (sometimes shortened to rate ratio) is the ratio of the incidence rate in those exposed to the incidence rate in those unexposed (see Box 2) (Lash 2021; Westreich 2019a). Many clinicians use "relative risk" when referring to the risk ratio. However, some epidemiologists use *relative risk* as an umbrella term for both risk and rate ratios, which can be a source of confusion when interpreting results (Lash 2021). For both risk and rate ratios, the null value when the risk is equal between groups is 1, and the measure can in theory take any value from 0 to infinity. Protective exposures will result in risk ratios or rate ratios less than 1, and harmful exposures will result in ratios greater than 1. Because these measures are on a ratio scale, results are typically communicated in multiplicative terms to avoid confusion with the risk difference (e.g., "the 5-year risk of myocardial infarction was 1.05 times higher in the exposed group"). The risk ratio, like the risk difference, should be interpreted in reference to the time of the study; this does not apply to the rate ratio.

Odds Ratio

Like risk and rate ratios, the odds ratio measures the association between exposed and unexposed groups, except that it is calculated using the incidence odds instead of the incidence proportion or incidence rate (Lash 2021; Westreich 2019a). Odds ratios are also interpreted the same way as risk and rate ratios, with a theoretical range of 0 to infinity, null value of 1, and harmful exposures resulting in odds ratios greater than 1, whereas protective exposures have odds ratios less than 1. However, in almost all cases that readers will encounter in the literature, the odds ratio will overestimate the risk of the outcome (Westreich 2019a) (i.e., for harmful exposures, the odds ratio will always be larger than the true risk ratio, and for protective exposures, the odds ratio will always be smaller [closer to 0] than the true risk ratio). If the outcome is uncommon (occurs in less than 5%–10% of patients), the odds ratio can be assumed to approximate the risk ratio; this is sometimes called the rare disease assumption.

The odds ratio will usually overestimate the risk of the outcome. Given this, it may be surprising that odds ratios are widely reported. There are two primary reasons for ongoing use of odds ratios (Westreich 2019a). The first reason is when other measures cannot be calculated. In some studies, data are insufficient to determine the total population at risk, either in number of individuals or time at risk; thus, incidence cannot be calculated. This situation is most common in case-control studies with cumulative control sampling, where odds ratios are typically the only measure of association that can be calculated and reported. The second reason is statistical. Specifically, logistic regression models report coefficients that are transformed to and reported as odds ratios.

Other Measures of Association

The earlier measures of association are some of the most commonly used in epidemiologic research and are common in the medical literature. Readers may encounter the incidence rate difference, prevalence difference, prevalence ratio, prevalence odds ratios, relative risk reduction, number needed to treat, and number needed to harm (Lash 2021; Westreich 2019a). The incidence rate difference and prevalence differences are analogs of the risk difference and are calculated using the incidence rate and prevalence proportion, respectively, instead of the incidence proportion. Similarly, the prevalence ratio and prevalence odds ratio are analogous to the risk ratio and odds ratio but are calculated with prevalence proportion and prevalence odds in place of incidence proportion and incidence odds, respectively (see Box 2).

Measures of association can be reported purely for descriptive purposes, but in the medical literature, they are often calculated with the goal of quantifying a potentially causal relationship (i.e., does the exposure [usually a treatment or intervention] cause an individual to move from one state to another with respect to a given outcome?) (Westreich 2019a). This is most evident in the number needed to treat and number needed to harm, which describe how treatment will cause change by preventing (or causing) occurrences of the outcomes. Despite the near-ubiquity of these causal questions in clinical research, they are rarely made explicit because many medical journals are reluctant to discuss causation in the context of study results (Saver 2019; Hernán 2018). This reluctance to make these causal questions explicit may partly be because of a lack of clarity regarding how to identify causal relationships between events, a fundamental question that has vexed philosophers and researchers for hundreds of years (Glass 2013). Over the past century, statisticians, epidemiologists, and others have developed or refined many causal frameworks to aid in making causal inferences, and although *causation* has historically been a term to avoid in medical research, the attitude toward causal inference may be shifting (Lash 2021; Lederer 2019). Next, we will consider new thinking about causal inference.

Causal Frameworks for Causal Inference

One of the most widely recognized causal frameworks is Sir Austin Bradford Hill's nine "viewpoints," which should be considered when trying to determine whether an observed association is causal (Box 3), commonly called the Bradford Hill criteria (Hill 1965). These criteria can be a useful framework for considering a potential causal relationship, but the relative simplicity of the Bradford Hill criteria has resulted in a

Box 3. Bradford Hill Considerations for Causation

- Analogy – similar cause-effect relationships exist among variables
- Biologic gradient – there is a dose-response or exposure-response pattern between cause and effect
- Coherence – the causal effect does not conflict with current scientific understanding
- Consistency – the association has been observed over time across studies with varying designs and populations
- Experimental evidence – removing or reducing the exposure decreases the frequency of the outcome
- Plausibility – there is a reasonable hypothesized mechanism that is consistent with a causal relationship
- Specificity – the cause has a single effect or an effect has a single cause
- Strength – the stronger the association, the more difficult it is to explain as a statistical artifact or bias
- Temporality – the cause precedes the effect

Information from: Hartung DM, Touchette D. Overview of clinical research design. Am J Health Syst Pharm 2009;66:398-408; Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295-300.

tendency to view them as a causality "checklist," even though the presence or absence of any single criterion cannot establish or rule out causality. Specific criticisms of each of the criteria have been described, and further efforts to refine the causal criteria have largely failed to generate consensus (Lash 2021; Holman 2001). Consequently, contemporary epidemiologic approaches have shifted away from causal criteria to focus on how causal relationships can be defined and how causal effects can be estimated. We discuss two widely used causal frameworks used in describing causal effects – the potential outcomes model and causal diagrams.

Potential Outcomes Model

The potential outcomes model, also called the counterfactual model, allows for causal relationships to be stated in clear mathematical terms and helps define the conditions that must be met, or assumed to be met, for causal inference to be justified (Lash 2021; Hernán 2020; Westreich 2019a). We are all well-versed in *counterfactual thinking*, even if we are not aware of it by that term. Counterfactual thinking is considering what would have occurred in alternative scenarios; in essence, these are "what if" questions, such as "what if I had ordered the steak instead of the fish?" In epidemiology, we may ask "what if the patient had not been exposed?" or "what if the patient had been treated?" Counterfactuals are alternatives to what occurred in reality; they are counter-to-fact. The potential outcomes refer to the different possible outcomes that could have occurred in these counterfactual scenarios.

The potential outcomes framework allows for causal effects to be clearly defined. Consider a patient who receives a heart transplant and dies 1 week later. Did the heart transplant cause her to die? This can be answered by considering the counterfactual question "what if the patient did not receive the transplant?" If she would still have died after 1 week, the heart transplant did not cause her death. If she would still be alive, the heart transplant did cause her death. This difference between factual and counterfactual outcomes is the individual causal effect. Readers have likely noticed that it is unknown which outcome would have occurred if the patient had not received the transplant, and until we develop the ability to time travel, we will never be able to go back, prevent this hypothetical patient's transplant, and observe what would have occurred. Indeed, because counterfactual outcomes are unobservable, individual causal effects in general cannot be estimated.

Potential outcomes and counterfactuals have thus allowed for a precise definition of an individual causal effect, but the effect cannot be measured. However, using this framework, average causal effects can be estimated from all the individual causal effects in a population under specific conditions for causal identification. These causal identification conditions include exchangeability, positivity, and consistency (Hernán 2020; Westreich 2019a). Exchangeability is the condition that the average baseline risk in the exposed group is equal to the average baseline risk among the unexposed. Essentially, in exchangeable populations, the groups receiving treatment and control could be changed and the results would still be the same. Positivity states that at baseline, all individuals could potentially receive any of the treatments in the study. Consistency means that the exposure causes the same effect in all individuals who are exposed. A consistent exposure is one that is defined clearly and specifically enough that, within the exposed and unexposed groups, any variations in the treatment received (e.g., dose, frequency, intensity) do not have different effects on the outcome. For example, an intervention of "1 aspirin per day" could mean 81 mg or 325 mg, and in a study of bleeding events, differences between these could result in a meaningful variation of outcomes between different doses. In such a case, the intervention "1 aspirin per day" would lack consistency. All of these conditions (exchangeability, positivity, consistency) presuppose that the exposure precedes the outcome and that all variables are measured accurately; thus, temporality and no measurement error are often included as additional causal identification conditions (Westreich 2019a).

When these conditions are true, the potential outcomes framework allows for the estimation of population (not individual) causal effects without needing to resort to time travel. Consider a hypothetical randomized trial population, where 10 people receive a transplant and 10 do not; all are followed for 1 week to assess mortality. A comparison of mortality between those who received the transplant and those who did not would provide a measure of association between the heart transplant and the outcome, and it can be calculated on the basis of observed (i.e., factual) outcomes only (no time

travel needed). If the causal identification conditions hold, according to the potential outcomes framework, the observed risk among those who received a transplant can stand in for the risk in the unexposed if they had, counter-to-fact, been exposed, and the observed risk among those who did not receive a transplant can do the same for the exposed if they had not been exposed. Thus, the comparison of observed risks between groups is equivalent to a comparison of counterfactual risks; therefore, the measure of association from the population is a measure of the true causal effect in the population, which is known as the average causal effect.

In summary, the potential outcomes framework establishes that meeting the causal identification conditions is sufficient to allow interpretation of measures of association as average causal effects. However, outside randomized controlled trials (RCTs), it is rarely possible to prove that a study meets all causal identification conditions. Instead, causal inferences assume that the conditions have been met. This can be problematic, particularly for exchangeability, because an individual's exposure status is often influenced by factors that also affect the risk of the outcome. Thus, the average risk in the exposed will differ from that in the unexposed, and exchangeability will not hold. However, if the factors that affect the risk of the outcome can be accounted for, conditional exchangeability can be achieved (i.e., groups are exchangeable conditional on these additional variables). This is sometimes called unconfoundedness or "no unmeasured confounding." The challenge, then, is to identify what these additional variables are. In the past 20 years, causal diagrams have emerged as a useful tool for addressing that challenge.

Causal Diagrams

Causal diagrams are directed acyclic graphs used for causal inference; they are a tool to visually codify the assumptions about the relationships between variables. Directed acyclic graphs are useful for identifying which variables affect exchangeability, as well as other potential sources of bias. Together with the potential outcomes framework, causal diagrams can guide the design, analysis, and interpretation of results, and readers evaluating a study can construct their own causal diagrams to see whether their assumptions align with the study methods. In this section, we provide a focused introduction to key elements of causal diagrams and how they can be used to identify sources of confounding and other biases. Interested readers can refer to any of the many published reviews for more information (Etminan 2020; Gaskell 2020; Lederer 2019; Shrier 2008).

In a causal diagram, variables are represented as nodes that are connected by arrows. Arrows represent direct causal effects between variables (e.g., E→Y indicates that E causes Y). A path is an unbroken sequence of variables connected by arrows, and these paths are the key to causal diagrams. If the arrows all point in the same direction (e.g., E→M→Y), it is a directed path; otherwise, it is undirected. A variable that comes before another variable on a directed path is an ancestor of the second variable, and the second variable is a descendent of the first (e.g., E→Y, E is an ancestor of Y, Y is a descendent of E). A causal path is any path between the exposure and the outcome under study where the arrows all point the same way (i.e., a directed path between exposure and outcome). In contrast, a non-causal path is any undirected path between exposure and outcome; these are the sources of bias when estimating causal effects.

Understanding how non-causal paths induce bias requires an explanation of mediator, collider, and common cause variables. A mediator is any variable that lies between exposure and outcome on a causal path (e.g., M in E→M→Y is a mediator). Mediators transmit causal effects from exposure to outcome. For example, sleep quality (exposure) can affect the outcome of work performance through the mediator of alertness (Figure 2A). A common cause is a variable with two arrows leaving it (e.g., C is a common cause of E and Y in the path E←C→Y [Figure 2B]). When the common cause is on a non-causal path, this variable is a confounder. For example, maternal age is a common cause of both birth order (exposure) and outcome of trisomy 21. A collider is a common effect of two other variables; it will have two arrows converging on it within a path (e.g., Z is a collider on the path E→Z←Y [Figure 2C]). For example, early in the pandemic, both occupation (exposure) and COVID-19 severity (outcome) may have influenced whether an individual was tested for COVID-19 (collider). Finally, in causal diagrams, a box around a variable is

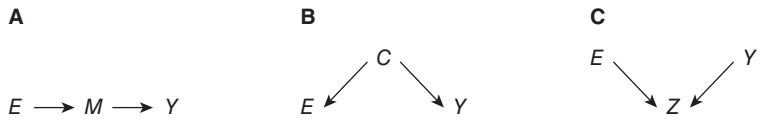


Figure 2. A, A causal path from E to Y through mediator M. B, A confounding path from E to Y through confounder C. C, A blocked non-causal path from E to Y. Z is a collider and, if conditioned on, would open the path, leading to bias.

Information from: Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020;158:S21-8; Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951-67; Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

often used to indicate it has been adjusted for ("conditioned on"). See Figure 2 for these variables and paths.

Colliders and confounders control whether non-causal paths induce bias. Non-causal paths can only induce bias if the path is open; blocked paths do not transmit bias. Confounding paths contain a common cause (i.e., a confounder) of the exposure and the outcome. Confounding paths are open by default and therefore cause bias. To block a confounding path, the confounder must be adjusted for. Of note, confounding paths can include many variables, including the common cause, and adjusting for any single variable is enough to block the path. For example, in the path E←C→D→Y, adjusting for either C or D is sufficient to block the path. In contrast to confounding, paths that contain a collider are by default blocked, and conditioning on the collider will cause "collider bias" or "collider stratification bias," a type of selection bias. Finally, adjusting for mediators also results in blocking paths. However, mediators lie on the causal path between exposure and outcome and generally should not be controlled for. Doing so will result in overadjustment, which generally reduces the magnitude of association between exposure and outcome, potentially obscuring the true causal effects (Lash 2021; Hernán 2020). Figure 3 provides hypothetical examples of common causal diagram structures, and Table 1 summarizes terminology and rules regarding these diagrams.

Although researchers often use causal diagrams to guide study design and analysis, pharmacists can also make use of these tools when evaluating studies. Using causal diagrams can identify potential sources of bias, such as open confounding paths or colliders that have been inappropriately conditioned on. To start, all the variables that a study controlled for in any way should be listed. Then, arrows should be drawn between variables to create a causal diagram that aligns with the pharmacist's subject matter knowledge. Once this diagram is constructed, variables can be identified as confounders, mediators, or colliders. If mediators or colliders were controlled for in the analysis, the study's analysis likely induced bias in the observed association. Figure 4 shows how this process can clarify a study's reported analysis.

A primary criticism of causal diagrams is that they are constructed on the basis of expert knowledge and theory, not data. This invites skepticism of causal diagrams as a tool because multiple causal diagrams could be proposed for the same causal relationship, and disagreements regarding which variables should be controlled for are possible. However, this can be a strength of causal diagrams. Drawing a diagram forces assumptions to be made explicit, allowing for collaborative discussion and critique. Discrepancies in causal diagrams are not errors, but signals of the need for further research in an area. Alternatives to causal approaches de-emphasize prior knowledge and purposeful deliberation in favor of probability and chance. In short, causal diagrams are useful in the never-ending pursuit for validity in epidemiologic research and better understanding of the world.

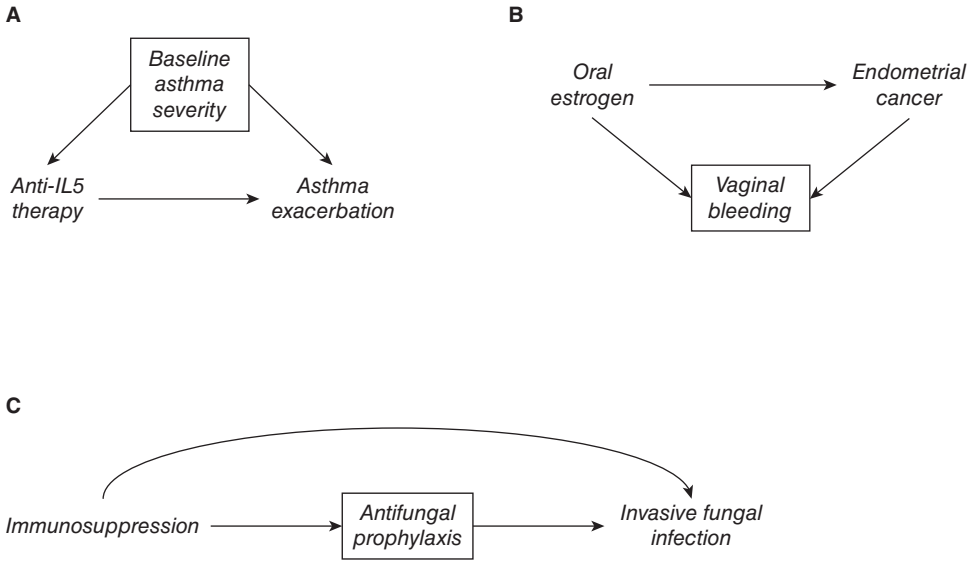


Figure 3. Hypothetical examples of causal diagrams. A, The effect of anti–interleukin 5 (anti–IL-5) therapy on asthma exacerbations is confounded by baseline asthma severity. The box around severity indicates it has been adjusted for, blocking the confounding effects. B, Oral estrogen causes endometrial cancer, but by conditioning on the collider of vaginal bleeding, collider bias (selection bias) occurs. C, Antifungal prophylaxis is a mediator of the effect of immunosuppression on invasive fungal infection. Adjusting for prophylaxis blocks the effect of immunosuppression on that pathway, causing overadjustment.

Table 1. Common Terminology and Rules in Causal Diagrams

Arrow	Indicator of direct causal effect between variables, the effect moves in the direction the arrow is pointing
Path	An unbroken sequence of variables connected by arrows.
Directed path	A path where all arrows point in the same direction
Undirected path	A path where not all arrows point in the same direction
Causal path	A directed path between exposure and outcome
Non-causal path	An undirected path between exposure and outcome
Ancestor	A variable that comes before a second variable on a directed path
Descendent	A variable that comes after a previous variable on a directed path
Mediator	A variable between exposure and outcome on a causal path
Collider	A variable on a non-causal that is a common effect of two other variables, it will have two arrows pointing in to it
Confounder	A variable on a non-causal path that is a common cause of two other variables, it will have two arrows pointing away from it
Confounding path	An non-causal path that contains a confounder of the exposure and outcome, allows for non-causal effects unless adjusted for
Blocked path	A path that contains noncollider (i.e., a mediator or confounder) that has been conditioned on, or a path that contains a collider that has not been conditioned on
Collider bias	Bias caused by conditioning on a collider on a non-causal path, this leads to unblocking of a blocked path
Overadjustment	Conditioning on a mediator
Box	A box around a variable indicates the variable has been conditioned on

Information from: Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020;158:S21-8; Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951-67; Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

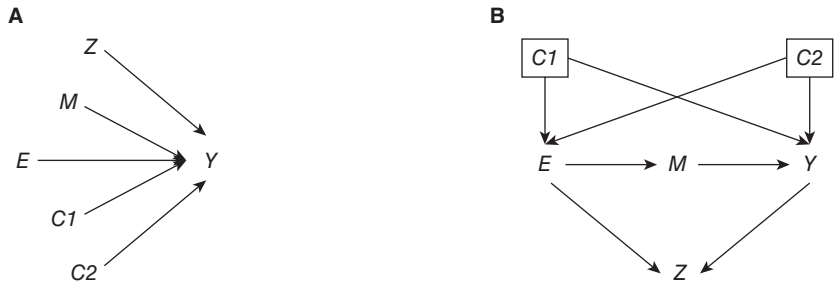


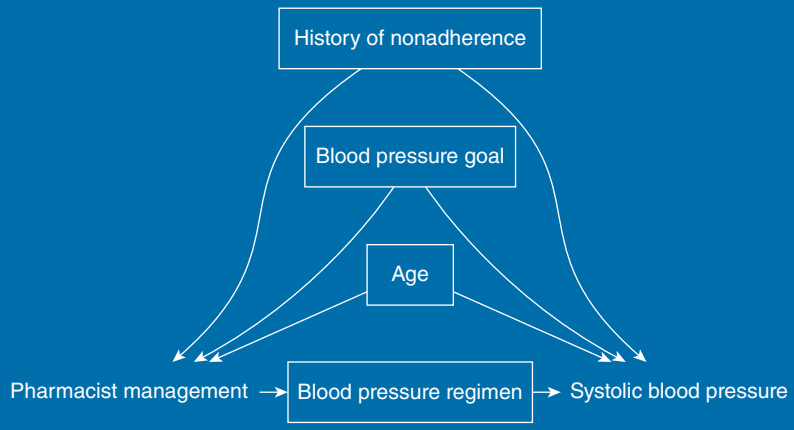
Figure 4. Causal diagrams can be a useful tool for critiquing published results of statistical models. Many investigators will include all measured variables in a model to be adjusted for, a situation that would appear like the causal diagram (A). However, with intentional consideration of the relationship between variables, an evaluator of a statistical model could identify that the causal diagram should look more like (B). With the knowledge of (B), it is apparent that only C1 and C2 needed to be adjusted for. Adjusting for M was overadjustment and Z induced collider bias.

Information from: Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020;158:S21-8; Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951-67; Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

Decision Scenario

A pharmacist is reviewing a cohort study of the pilot phase of a new pharmacist-run hypertension management service, where the pharmacist may adjust medication regimens if needed. In this study, clinic providers could choose to refer patients to the new pharmacist-run service or the existing nursing-led blood pressure monitoring program. To evaluate efficacy of the new service, the investigators used linear regression to compare the mean systolic blood pressure between groups after 3 months of follow-up. To adjust for confounding, the authors included patient age, blood pressure regimen during the 3 month follow-up, blood pressure goal, and history of nonadherence.

To evaluate the methods, the pharmacist reviewing the study constructs the following causal diagram with boxes around the variables that have been adjusted for.



Assuming the diagram is correct, what conclusions can the pharmacist make about the validity of the study results?

ANSWER

Causal diagrams are extremely useful for understanding the relationships between variables, and drawing a diagram can help evaluate the appropriateness of the data analysis performed in a study. In this example, the exposure is referral to the pharmacist-run hypertension service, and the outcome is mean systolic blood pressure at after 3 months of follow-up. The regression model adjusted for age, blood pressure goal, blood pressure regimen during follow-up, and history of nonadherence. In the corresponding causal diagram, age, blood pressure goal, and history of non-adherence are identified as plausible confounders. These variables could influence if a patient gets referred to the pharmacist service or not, and also the extent of blood pressure reduction observed. Controlling for confounders blocks confounding, allowing for estimation of causal effects, and therefore it was appropriate to include these variables in the model. In contrast, blood pressure regimen is identified as a mediator variable in the causal pathway (pharmacist management → blood pressure regimen → systolic blood pressure). This is plausible, as the pharmacist running the service can adjust medication regimens, and different medication regimens might affect systolic blood pressure to different degrees. Controlling for a mediator blocks the causal pathway, causing overadjustment, so mediators generally should not be controlled for. Therefore, assuming the proposed causal diagram is correct, the results of the study will likely be biased due to overadjustment from controlling for the mediator blood pressure regimen.

1. Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020;158:S21–8.
2. Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951–67.
3. Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

Validity in Epidemiologic Research

Validity can be thought of as a consideration of how well a measure describes the phenomena it is intended to describe (Hulley 2013). In epidemiologic contexts, study validity is "the degree to which the inferences drawn from a study are warranted when account is taken of the study methods and the characteristics of the participants in the study" (Porta 2008) (i.e., how closely do the results of a study reflect the underlying truth of the relationship between exposure and outcome?). The goal of high-quality epidemiologic research is to produce valid estimates of the relationship between an exposure and an outcome.

Validity is often further divided into internal and external validity (Westreich 2019a; Hartung 2009; Porta 2008). Internal validity refers to the extent that the measured effect in the study sample matches the true effect in the study sample (Westreich 2019a). The major threats to internal validity are random error (chance), systemic error (bias), and confounding

(Westreich 2019a). Random error is generally reduced through increased sample size, and the amount of random error in a study result can be quantified through appropriate statistical tests (Vetter 2017; Rothman 2012; Hartung 2009). In contrast, managing bias and confounding requires more careful consideration in study design and more complex statistical approaches than most simple inferential tests. External validity, in contrast, is the degree to which the true effect in the study sample matches the true effect in the broader population of interest (Lash 2021; Westreich 2019a). External validity considers the extent to which a study's results can be generalized to the population the study sample was obtained from (generalizability) and transported to other, different populations that were not sampled as part of the study (transportability). Internal validity is often considered a prerequisite of external validity, which is true to a point (Delgado-Rodriguez 2004). However, many design choices that enhance internal validity may compromise external validity, and some epidemiologists have advanced "target validity" as an approach to balance internal and external validity, depending on the goal of the study (Westreich 2019b). The focus of subsequent sections of this chapter is on internal validity, though we will occasionally discuss considerations of external validity.

Bias

Bias arises when there is a systematic error in the design or conduct of a study that results in nonrandom deviation of a study's estimate of effect from the true effect, threatening internal validity (Vetter 2017; Hulley 2013; Gerhard 2008; Porta 2008). At least 70 types of bias encountered in epidemiologic research have been described, and many attempts to categorize and organize them have been proposed (Delgado-Rodriguez 2004; Maclure 2001). Most commonly, biases are grouped into three categories: selection bias, information bias, and confounding. We will discuss selection bias and information bias, including common types of bias in each category. We will also discuss time-dependent biases, which are of particular concern in pharmacoepidemiology and can result in either information bias or selection bias. Confounding is discussed separately in the text that follows.

Information Bias

Measurement is at the center of epidemiologic studies. At a minimum, an epidemiologic study requires measurement of the outcome in question; however, most studies also evaluate the effects of an exposure, which also need to be quantified. In addition, other variables, such as potential confounders or mediators, are typically measured. Each time a variable is measured, there is a potential for error in that measurement. Information bias occurs when there are systematic errors in how information is measured in a study. Misclassification is a measurement error that occurs with categorical variables and results in an individual's placement in the incorrect category, such as classifying a smoker as a nonsmoker.

When misclassification occurs equally across all groups in the study (e.g., exposed vs. unexposed, with outcome vs. without outcome), non-differential misclassification occurs. Non-differential misclassification typically arises from decisions made in operationalizing a study, such as using measurement tools or categorization rules that are inaccurate or non-validated, or even failing to clearly define what constitutes an exposure or outcome event. As a clinically relevant example, consider diagnosis codes. A study of the accuracy of the International Classification of Diseases, 10th Revision (ICD-10) diagnosis codes to identify UTIs found that, compared with chart review, diagnosis codes accurately classified UTIs as present or absent only about 63% of the time (Livorsi 2018). If an observational study used ICD-10 codes to define the outcome for all patients, it would misclassify the outcome almost 40% of the time. Because the outcome would be defined in the same manner for all patients, this misclassification would be non-differential.

Conversely, when misclassification instead occurs unequally across individuals in a study, the result is differential misclassification. Differential misclassification typically occurs when exposure status influences how outcomes are measured, or when knowledge of the outcome affects how exposure status is captured. One of the most common types of differential misclassification is recall bias. Recall bias occurs when patients are asked to recall exposures after the outcome has already occurred and knowledge of the outcome affects the individual's ability to recall exposure data. The classic example of recall bias is in studies of congenital disorders, where it is thought that because of the adverse pregnancy outcome, parents of newborns with a congenital condition are more likely to correctly recall potential exposures that could have caused the disorder (Rothman 2012). Parents of healthy newborns lack such a stimulus; thus, they are not expected to recall exposures with the same effort. The result is that an exposure is more likely to be accurately classified for subjects with congenital disorders, resulting in differential misclassification of exposure status.

Although the distinction between differential and non-differential misclassifications may seem academic, the consequences of these different types of bias are of practical importance. When non-differential misclassification occurs while categorizing a binary variable (e.g., exposed [yes/no], experienced outcome [yes/no]), the result will always be an estimated measure of association that is biased toward finding no difference ("biased toward the null") because errors in classification of exposure and/or outcome lead to dilution of the true effect. This is one of the most common forms of information bias in the medical literature, and it is important to recognize that it results in an underestimate of the true effect. Non-differential misclassification of a binary variable should therefore be of particular consideration in studies that find no difference, especially if the outcome under investigation is harmful. In contrast, differential misclassification

(and non-differential misclassification of a multicategory variable) can lead to estimates that are smaller than the true effect (biased toward the null) or amplification of the estimated effect compared with the true association (biased away from the null). The direction of bias is difficult to predict and will depend on the structure of the relationship between variables.

All epidemiologic studies are at risk of information bias because all measurements will have some error (Lash 2021). Investigators should focus on designing studies that minimize the magnitude of measurement error and misclassification. The specific approaches will depend on the methods of data collection and the potential source of bias being addressed. For example, recall bias could be minimized by framing questions to improve all subjects' ability to recall information, or by using information documented in the medical record, thereby avoiding the need for any recalled information at all. For many variables, relying on documented information will be more accurate than relying on reported measures. Validated questionnaires, interviewer training, blinding, and standardized questions can all improve the collection of subjective data or information that cannot otherwise be obtained from documented records. Retrospective studies may be more susceptible to information bias because of lack of investigator control regarding when and how data were collected. In addition to mitigating measurement error through study design, statistical approaches have been developed that try to correct measurement error or quantify the potential magnitude of bias (sensitivity analysis) (Lash 2021). These bias analyses are seldom reported because they can be technically complex and typically require additional information or assumptions about the structure and magnitude of the error.

Selection Bias

The terminology related to selection bias is highly inconsistent, and many different disciplines use the term *selection bias* in different ways, with resulting differences in interpretation. In econometric and medical literature, selection bias typically refers to "treatment selection bias" – systematic differences in why someone receives one treatment or another. Epidemiologists consider this a type of confounding. In epidemiology, selection bias refers to biases that occur as a result of how patients are selected into the study (not how they self-select or are selected to different treatments) (Lash 2021). The result of selection bias is that the estimated measure of effect obtained from the study population will differ systematically from the true association in the source population (Lash 2021). Selection bias is difficult to detect objectively because data regarding the association in the source population are usually unavailable; thus, the presence of selection bias must be inferred (Rothman 2012).

Selection bias, in the epidemiologic sense, generally arises from three main sources. The first source of selection bias comes from how patients are selected into the study population. However, study entry is only the beginning of a patient's participation, and not all patients remain in the study for the entire follow-up. The second source of selection bias arises during follow-up if patients drop out of a study for reasons related to the exposure and outcome. Loss to follow-up might not seem to align with the idea of selection, but it can be conceptualized as selecting only patients with complete follow-up. The third source of selection bias – that arising from restricting to or adjusting on a collider variable – can be the most difficult to detect because it occurs during data analysis. This last source is conceptually challenging because it is difficult to intuit the connection between data analysis and selection into the study. Nevertheless, it has clearly been shown that colliders are a consistent source of selection bias (Hernán 2004).

Depending on the relationship between the exposure, outcome, and factors influencing selection into the study, selection bias can take on two structural forms (Hernán 2017, 2004). The first form occurs when selection (or retention) into the study is a collider (i.e., the result of factors related to both exposure and the outcome). As a clinical example, consider a trial of calcium channel blockers compared with other antihypertensives on the risk of liver injury. If calcium channel blockers cause more adverse effects, more patients in that treatment group may drop out. Separately, patients with alcohol use disorder may be more likely to drop out as well as more likely to experience the outcome. Colliders induce bias because knowledge of the collider and treatment provides information about the outcome. If we know someone did not drop out and they received a calcium channel blocker, they are less likely to have alcohol use disorder and therefore liver failure. This makes calcium channel blockers appear protective, even if they truly have no effect; this example can be shown in a causal diagram (Figure 5). This form of selection bias occurs when a study restricts or stratifies on a variable that is a collider. This type of selection bias can also occur if a researcher statistically adjusts for ("conditions on") a collider, which can be a source of confusion because

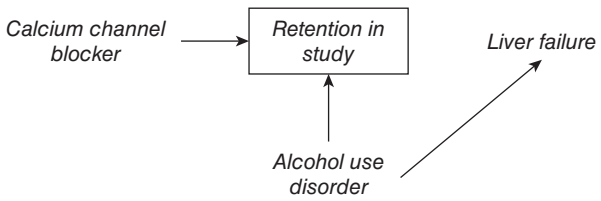


Figure 5. Selection bias in a randomized controlled trial. Retention in study is a collider of treatment and alcohol use disorder. Loss to follow-up could therefore cause selection bias, even in this hypothetical randomized controlled trial.

statistical analysis is often thought of as separate from study selection procedures. Therefore, causal diagrams are particularly effective at detecting potential selection bias. No matter where the collider bias occurs (e.g., sample selection or analysis), the result is a biased estimate of the relationship between the exposure and the outcome in the study population and lack of generalizability to the source population.

The second form of selection bias does not involve collider variables. This form of selection bias occurs when selection into the study is associated with the outcome, but only in settings where the exposure has a true effect on the outcome. Envision a study of a screening test for colon cancer (Rothman 2012). Investigators randomly assign participants to screening or no screening and collect data on the incidence rate of colon cancer. The screening test should detect more colon cancer and thus affect colon cancer rates. However, patients who volunteer for such a trial may do so because of a family history of colon cancer; these individuals would have a higher risk of the outcome. In this hypothetical study population, therefore, the rate of colon cancer will be higher than in the source population, and the resulting estimated measure of effect for screening will therefore be a biased estimate of the true effect in the source population (which includes people without a family history of colon cancer). However, the estimate would be an unbiased estimate of the effect of screening within the study population because the selection bias was equally distributed between the randomized groups. Therefore, the consequence of selection bias without colliders is related to the impact on the generalizability of the study results, rather than the results being biased within the context of the study itself.

Because every study involves selecting patients for inclusion from a source population, every study is at risk of selection bias. Consequently, minimizing selection bias starts with sound study design. Selection bias can almost never be addressed through common adjustment methods such as regression. In fact, as we saw, selection bias can even be caused by inappropriate adjustment for a collider variable. Causal diagrams that are based on expert knowledge related to the research questions are one of the best tools to prevent or diagnose potential selection bias. Collider variables should not be used as criteria for selection into a study, nor should they be adjusted for in an analysis. Loss to follow-up may also lead to selection bias, and when possible, efforts should be made to obtain complete follow-up data from all subjects, which is easiest in prospective studies. As with information bias, quantitative bias analysis methods have been developed that can help identify the presence and potentially the magnitude of selection bias. However, these methods require additional data about the source population that are rarely available or assumptions about selection probabilities that can be difficult to justify. Consequently, selection bias analysis is not commonly performed in observational studies.

Time-Dependent Biases

Many time-related biases have been described, with immortal time bias and time-window bias of particular note (Suissa 2020). Immortal time bias can occur in cohort studies when there is a window of time between when a patient enters a cohort and when treatment is initiated (Suissa 2008). This period between cohort entry and treatment initiation is called "immortal time" because the patient must survive (or more generally, remain in the cohort) without experiencing the outcome long enough to receive the treatment. More generally, the bias occurs when treatment status is determined by events that occur after initial cohort entry. During the immortal period, the subject has not actually received treatment yet; bias occurs when a patient's immortal time is retroactively categorized as "treated" (misclassification) or excluded from analysis (selection bias). For example, consider a study of statin use to prevent myocardial infarction, where "statin use" is defined as three prescriptions filled within 6 months of cohort entry. Patients become eligible for the study at the time of hospital discharge after cardiac catheterization. Immortal time bias occurs because all patients who are "treated" had to survive long enough to fill three prescriptions. Patients who experienced a myocardial infarction after filling one or two statin prescriptions would be classified as a nonusers by this study definition because they had not yet filled three prescriptions, which is ultimately misclassification (information bias). Alternatively, excluding such a patient from the analysis would result in selection bias.

Time-window bias is a time-dependent bias that occurs in case-control studies. Time-window bias occurs when the exposure time-window differs between cases and controls. For cases in a case-control study, exposure must occur before onset of the outcome; however, with cumulative control sampling, exposure among controls can occur at any point during the study. A control could theoretically be "exposed" on the last day of follow-up. Thus, relative to cases, controls have a longer "time window" over which to be exposed, and later exposures decrease the time during which the outcome can occur, which can lead to time-window bias (Suissa 2020, 2011).

Both immortal time bias and time-window bias can be avoided with proper study design. Because the immortal time bias arises when cohort entry and treatment status are not aligned, the ideal solution is to design a study in which treatment is initiated at the same time as the patient becomes eligible to be included in the study, much as would be expected in a clinical trial. However, there are often limitations that make it difficult or impossible to design such a study. In these cases, one of the most common tools to address potential immortal time bias is to include treatment as a time-varying covariate in a statistical model. In time-varying models, patients can be categorized as untreated for the study period before they begin treatment and then change to the treatment group thereafter. This approach can usually help mitigate

immortal time bias. Many other approaches have also been suggested, but they are often less effective at reducing bias or are technically complicated. For time-window bias, sampling controls with the same duration of exposure as cases can minimize the bias; this is a sampling approach known as incidence-density sampling.

Confounding

Confounding is a widely recognized problem in observational studies that many pharmacists will be familiar with. Often described as a "mixing of effects," confounding occurs when an apparent association between an exposure and an outcome is caused by a third variable. Formally, confounding is bias from a variable that is a common cause of the exposure and outcome. Although the concept of confounding is fairly well understood, identifying the confounders is not always straightforward. Traditionally, a confounder has been defined as any variable that is (1) associated with the exposure, (2) associated with the outcome in the unexposed, and (3) not an effect of the exposure (Hernán 2020; Rothman 2012). However, contemporary work has shown that this definition can describe colliders as well as confounders (Hernán 2020). Causal diagrams have helped address these challenges and clarified the definition of confounding and confounders. Confounding is a non-causal path created by a common cause of the exposure and outcome. A confounder is therefore any common cause variable on the confounding path (Westreich 2019a).

Effect measure modification is a concept that is distinct from, but often related to, confounding. Confounding normally has the same effect on the exposure-outcome relationship for all individuals. However, sometimes, the observed measure of effect varies across levels of another variable; this is effect measure modification. For example, if a study of calcium supplementation and fractures reports a risk ratio of 4.0 for women and 0.2 for men, effect measure modification by sex is likely present. Effect measure modification is identifying a subpopulation that is particularly susceptible to the effects of the exposure. Conventional recommendations are to report the level-specific measures of effect rather than a single population estimate; however, sometimes, a population estimate may be preferred (Westreich 2019a).

Confounding is expected in all observational studies. Similar to selection bias and information bias, confounding can be addressed during the design phase using strategies such as restriction or matching. As noted earlier, confounding can also be addressed during the analysis phase through stratification or many other statistical approaches. We address these tools in greater detail in the sections that follow.

EPIDEMIOLOGIST'S TOOLBOX I: VALIDITY IN STUDY DESIGN

In the quest for high-quality research that moves the profession forward and closer to understanding the unbiased and potentially causal relationship between exposures and outcomes, it is essential that researchers and those who evaluate the medical literature understand the tools available in the epidemiologist's toolbox for both study design and analysis.

Addressing Bias Through Study Design

Sound principles of study design are the foundation for conducting valid and meaningful research. Choices made during the design phase influence the potential for confounding and are instrumental in minimizing many types of bias. Although bias can sometimes be adjusted for during the analysis, doing so often requires assumptions that are untestable. Thus, design is an optimal way to address bias and reduce the need for assumptions and complex analytical tools. It is assumed that readers are generally familiar with the basic elements of common analytical study designs. In this section, we review these study designs with particular consideration toward causal inference and potential sources of bias. Specific strategies for addressing bias were previously addressed.

Randomized controlled trials are considered the gold standard for establishing causal relationships. This is clear when considering the causal identification conditions. Randomization and treatment assignment provide exchangeability and positivity because treatment assignment is independent of patients' risk of the outcome, and all patients could potentially receive the treatment. The prospective design ensures temporality, and the assignment of a clear, well-defined treatment provides consistency (Hernán 2020). Consistency may be less certain in pragmatic trials, which usually have less restrictive enrollment criteria and allow more ancillary treatments; the results are typically more generalizable (Sedgwick 2014). Despite being the gold standard for causal inference, RCTs are not immune to bias. Measurement error, and thus information bias, is possible in any study, though generally, investigators of RCTs go to great lengths to ensure that measurements collected are meaningful, accurate, and valid. Nevertheless, consideration of the appropriateness of chosen study measures is essential to evaluating the risk of bias in an RCT (Sterne 2019). Objective measurements are less susceptible than subjective outcomes to differential measurement error. Clinical trials have the advantage of blinding participants and researchers to treatment assignment, which helps minimize the risk of differential assessment of subjective outcomes. Although selection bias is uncommon, it may still occur if there is extensive loss to follow-up that is related to both the intervention and the outcome. Selection bias can also occur if allocation is not concealed. In such a trial, researchers may choose whether or not to enroll a patient in a study at all, depending on knowledge of what treatment will be assigned and prognostic factors, leading to selection bias (Mansournia 2017). Immortal time bias should not occur in intention-to-treat analysis of RCTs because cohort entry and treatment assignment both occur at the time of randomization.

Cohort studies lack randomization; hence, exchangeability must be assumed conditional on measured confounders; this is a principle reason why claims of causality based on cohort studies are often viewed with skepticism. Without treatment assignment, positivity must also be assumed, and consistency may be difficult to ascertain. Care must be taken to define the exposure clearly and in such a way that it can be accurately measured. This can be particularly challenging with retrospective data. Cohort studies can generally establish temporality because individuals are followed forward in time from the exposure; however, determining the exact timing can sometimes be challenging, particularly in retrospective cohort studies. Bias is a problem that must be considered in every cohort study. Like RCTs, cohort studies are susceptible to information bias from the same potential sources of measurement error, and the risk is likely increased in cohort studies, particularly retrospective cohort studies. In such studies, investigators have little control over how data are collected, and measurements of variables may be of varying quality and accuracy, assuming the data are present at all. Furthermore, blinding of participants is not possible in cohort studies, creating opportunities for differential misclassification. The potential for selection bias is also increased in cohort studies. Like in RCTs, selection bias may occur in cohort studies when loss to follow-up is related to both the exposure and the outcome. However, in cohort studies, efforts to track patients over time may be less rigorous (or absent). Selection bias can also occur when defining the population of a cohort study, for which Table 2 provides an example. Finally, lack of treatment assignment places cohort studies at risk of immortal time bias because exposure status may be determined on the basis of information collected after cohort entry.

Case-control studies have many of the same challenges as cohort studies with respect to causal inference and bias. Exchangeability must be conditional on measured confounders, and positivity must be assumed. Consistency is also difficult to ascertain. Temporality is particularly challenging in case-control studies because outcome status is determined first, and care must be taken when looking back to ensure the outcome was not present when the exposure occurred. The design of case-control studies also makes interpretation difficult. Most case-control studies use cumulative control sampling. With this sampling method, cases include all individuals with the disease, but controls represent only a sample of the population who never developed the disease by the end of the study. Consequently, the total population at risk of the outcome during the study is unknown. Therefore, incidence proportions (risk) cannot be calculated, and a risk ratio calculated from a case-control study will be a biased estimate of the true risk ratio in the total population. Instead, incidence odds and odds ratios must be calculated. Alternative control sampling strategies have been developed that can allow for estimation of risk ratio or incidence rate ratio, but these are less common in the medical literature (Westreich 2019a; Rothman 2012). Case-control studies are susceptible to many of the same biases as other studies and, in some instances, more so than other designs. Selection bias, in particular, is of primary concern in case-control studies. Sampling of controls should be done without regard to exposure status, but sometimes, this is not the case. Strategies to make control identification easier, such as recruiting friends, family, or neighbors, may lead to biased sampling. Similarly, recruiting only patients from the hospital or those treated by certain physicians can cause selection bias. Table 3 and Figure 6 illustrate selection bias in a case-control study. Case-control studies are also vulnerable to information bias, and differential misclassification as a result of biases such as recall bias is of particular concern. Time-window bias may also be a concern in case-control studies (Suissa 2011).

Table 2. Selection Bias in a Cohort Study

Study Design	Retrospective Cohort Study
Population	Women with recent vaginal bleeding
Exposure	Already on oral estrogen
Control	Not on oral estrogen therapy
Outcome	Endometrial cancer
The bias:	Both oral estrogen therapy and endometrial cancer can cause vaginal bleeding. A woman on estrogen therapy may be more likely to have experienced estrogen-induced bleeding, whereas vaginal bleeding in a woman not on estrogen is more likely to be because of endometrial cancer. Thus, by selecting only women with recent vaginal bleeding, the researchers would have induced a negative association between the two (vaginal bleeding in a woman not on estrogen means she is more likely to have cancer and vice versa). This bias is illustrated in Figure 3B

Information from: Greenland S, Neutra R. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. J Chron Dis 1981;34:433-8; Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12:313-20.

Addressing Confounding Through Design

As described earlier, investigators of observational studies face many hurdles to causal inference, and assumptions must be made regarding all the causal identification conditions. One of the primary issues is lack of exchangeability because of bias and confounding. Previous sections have identified

Table 3. Selection Bias in a Case-Control Study

Study Design	Retrospective Case-control Study
Population	Cases were individuals hospitalized with pancreatic cancer; controls were selected from patients being treated by the same physicians as the cases
Exposure	Coffee consumption
Control	No coffee consumption
Outcome	Pancreatic cancer
The bias	In this study, cases were patients hospitalized with pancreatic cancer, and controls were selected from hospitalized patients seen by the same physicians (many who were likely gastroenterologists) but without pancreatic cancer. Many of these controls had other GI diseases and possibly had decreased coffee intake (either because of physician recommendation or because of their own volition). Sampling of controls was thus not independent of exposure, and coffee consumption among controls was not representative of the total population, biasing the study results. This bias is illustrated in Figure 6

Information from: MacMahon B, Yen S, Trichopoulous D, et al. Coffee and cancer of the pancreas. N Engl J Med 1981;304:630-3.

several potential sources of bias and how studies might mitigate these biases; in this section, we consider strategies for ensuring exchangeability. Assuming that other causal identification conditions hold, achieving conditional exchangeability may allow for causal inferences from observational data.

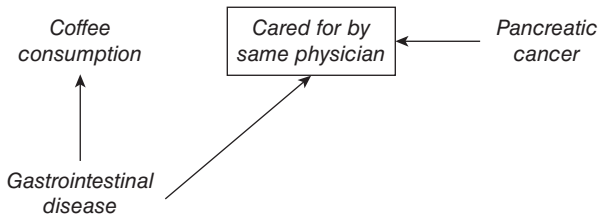


Figure 6. Selection bias in a case control study (see Table 3 for details).

Information from: MacMahon B, Yen S, Trichopoulous D, et al. Coffee and cancer of the pancreas. N Engl J Med 1981;304:630-3.

Randomization

Randomization is the most effective design tool for addressing confounding, particularly because it addresses both measured and unmeasured confounders. The benefits of randomization in providing exchangeability were previously discussed, and in this section, we emphasize a few specific points to consider when evaluating RCTs. The first is that randomization only eliminates confounding for intention-to-treat analyses. Intention-to-treat analyses evaluate patients according to the group to which they are assigned and are thus unconfounded. However, intention-to-treat analyses answer the question "what is the effect of assignment to treatment X?" Often, the desired question is "what is the effect of receiving treatment X?", which is addressed by per-protocol analyses. Per-protocol analyses may still be confounded, given that whether or not a patient actually receives treatment may be influenced by many variables. In this regard, per-protocol analyses are sometimes regarded as cohort studies nested inside RCTs (Westreich 2019a). In addition, it is important to remember that randomization only addresses confounding at baseline; problems that occur during the study, such as loss to follow-up, can still result in confounding and selection bias. Finally, because randomization works by balancing the average risk of outcome between treatment groups, it can sometimes "fail" for studies with small sample sizes, resulting in residual imbalances between groups. Evaluation of baseline characteristics for imbalances can signal potential problems with the randomization process (Sterne 2019).

Restriction

However, randomization is often not possible, and most epidemiologic studies must rely on other design and analytical approaches to address confounding. One of the simplest approaches is to limit the study population to only individuals with or without the suspected confounder, an approach known as restriction. For example, researchers concerned that a new air freshener is associated with lung cancer may choose to include only individuals who have never smoked. Restriction eliminates confounding for the variable in question because, put simply, "a variable cannot produce confounding if it is prevented from varying" (Lash 2021). For continuous variables such as age, studies may age restrict to a narrow age range (e.g., patients 45–55 years of age). In these cases, confounding from the continuous variable will be reduced but not eliminated.

Although restriction is a simple approach that is relatively easy to implement, it has several limitations. On a practical level, it is only possible to restrict on variables that are measured; thus, restriction cannot address unmeasured confounding. In addition, restriction reduces the study sample size; thus, the number of restrictions will be determined in part by the number of potential subjects. Restriction can be combined with other methods for addressing confounding, which can overcome some of these practical barriers.

Limiting the study sample to certain groups also decreases generalizability because the sample becomes less representative of the target population, though some epidemiologists have argued that this tradeoff is acceptable if the restriction allows more accurate estimates of the measure of effect (Lash 2021). One of the most important considerations with restriction is to ensure that the restricting variable is a confounder, and not a collider or mediator. As discussed, if the sample is restricted based on a collider, this will cause selection bias. This was the concern with the proposal to restrict a study of estrogen therapy and endometrial cancer to women with recent vaginal bleeding (see Table 2).

The limitations of restriction should be considered when evaluating a study that uses this approach. Pharmacists appraising such studies should assess how completely the restrictions addressed confounding from those variables, what confounders were not addressed by restriction, the methods used to address the remaining confounders, and the risks of unmeasured confounding. In addition, the risk of selection bias from restricting on a collider should carefully be considered by examining the structure of the relationships between the restricting variable, the exposure, and the outcome.

Matching

With matching, the goal is to create pairs or groups of subjects that are similar to each other except for exposure status. For example, in a study evaluating the risk of acute kidney injury from vancomycin plus piperacillin/tazobactam compared with vancomycin plus cefepime, each patient with septic shock in the piperacillin/tazobactam group was matched to a patient with septic shock who received cefepime (Navalkele 2017). Unlike restriction, matching allows for inclusion of people with and without the confounder. Matching controls for confounding from each matched variable, and as with restriction, more than one variable can be matched on. For continuous variables, matching may be either to specific values (e.g., 25 years old) or to a range of values (e.g., 45–55 years of age), and the potential for residual confounding should be considered.

Matching is generally recommended because it can increase statistical precision and is thus seen as an "efficient" tool. However, matching can take additional time and financial resources and may not be financially "efficient." Matching has many of the same limitations as restriction. Like restriction, matching can only occur for variables that have been measured, and residual confounding from unmeasured confounders is likely. Increasing the number of variables matched on decreases the probability of finding a match for any given patient, and unmatched subjects will be excluded from the study, reducing the total sample. When the initial sample pool available for matching is small, matching is typically only feasible for a few confounders. Clinicians evaluating studies that use matching should keep these limitations in mind and consider the potential for remaining confounding and how the study addressed it. In addition, the analysis of matched data is full of pitfalls that pharmacists should be aware of; these are discussed in a later section.

EPIDEMIOLOGIST'S TOOLBOX II: VALIDITY IN DATA ANALYSIS

In the absence of randomization, design choices alone are often insufficient to address confounding. Additional analytical approaches are often used to try to adjust for confounding, with the goal of achieving conditional exchangeability given the measured confounders. Except for instrumental variables, none of the following analytical tools can address unmeasured confounding. As in the previous section, our discussion in the section will assume that study design has minimized other biases and that assumptions related to other causal identification conditions are reasonable.

Stratification

One of the best-established methods of controlling for confounding is stratification, which has been described as the mainstay of epidemiologic analyses when consideration of factors beyond the exposure and outcome (i.e., confounders) is required (Lash 2021). In stratified analyses, subjects are grouped according to levels of the confounder, and for each level, a separate measure of effect is calculated. For example, if smoking status were considered a confounder of the relationship between statin use and myocardial infarction, the data would be stratified according to smoking status (e.g., "current," "former," and "never" smokers), and within each stratum, the risk ratio (or other measure of effect) for myocardial infarction in patients treated or not treated with statins would be calculated.

In its most basic form, the result of a stratified analysis results in multiple stratum-specific measures of effect, which are particularly beneficial if looking to assess or report effect measure modification. However, in many cases, a single overall estimate of the association between exposure and outcome is desired. Stratum-specific estimates can be pooled together to produce such an estimate. One of the most common methods for combining stratum-specific estimates is the Mantel-Haenszel (sometimes Cochran-Mantel-Haenszel) approach, which returns the average of the stratum-specific estimates, weighted proportionally to the number of individuals in each stratum. The Mantel-Haenszel method assumes that the effect of the exposure is the same (or very similar) for all levels of the confounder (i.e., no effect measure modification). An alternative approach to combining strata that does not make that assumption is standardization. Standardization also weights the number of events in each stratum, but the weights are typically generated from an external population reference. For example, if the hypothetical study of statins and myocardial infarction stratified by smoking took

place in the United Kingdom, the estimates in each level of smoking would be weighted by the proportion of current, former, and never smokers in the total population of the United Kingdom.

Stratification is quite similar to restriction, but instead of excluding certain groups, stratification allows all patients to be retained in the study. The relatedness of these two methods means that many of the limitations of stratification will resemble those of restriction. Like restriction, stratification is relatively simple and easy to implement and can be quite useful for a small number of confounders. However, stratification quickly becomes untenable when multiple variables are used. If investigators of the earlier study of statins and myocardial infarction wished to stratify on biological sex (male or female) in addition to smoking status, six strata would be created (three levels for smoking and two levels for sex). Adding age in five categories would result in 30 (3 × 2 × 5) different smoking-sex-age strata. With this many levels, many of the strata would likely have few or no patients. Another important consideration with stratification is the choice of stratifying variables. Recall that restricting on a collider induces selection bias; stratification on a collider has the same effect. In fact, this type of selection bias is sometimes called collider-stratification bias.

Regression Modeling

Regardless of whether they realize it, most pharmacists are familiar with the concept of regression modeling. In its simplest form, regression modeling can be conceptualized as drawing the line of best fit through some data (think "$y = mx + b$" from algebra), though in practice, regression modeling becomes much more complicated when applied to epidemiologic research (Grant 2019). With increases in computing power over the past decades, regression has emerged as one of the most widely used statistical tools to control for confounders in observational studies. The outcome variable in the regression model is sometimes called the dependent variable. The exposure and confounders, though independent variables, are often called predictors or covariates. A univariable regression model consists of one outcome variable and one predictor, whereas a multivariable model includes one outcome variable but more than one predictor. In a multivariate model, multiple outcomes are being modeled simultaneously; often, when studies report multivariate modeling, they are referring to multivariable modeling.

Pharmacists reading epidemiologic or medical literature are likely to come across linear, logistic, Poisson, and Cox proportional hazards regression models. The choice of model depends on the nature of the outcome being modeled. Linear regression is used when the outcome variable is measured on a continuous scale, such as systolic blood pressure or A1C. Count data, such as the number of asthma exacerbations or the number of goals scored in a soccer match, often look like continuous data. However, count data are always nonnegative integers and are often skewed. Poisson regression is often used for count data. Logistic regression is appropriate for binary outcomes, such as 30-day mortality or clinical cure. Finally, time-to-event (survival) outcomes, such as time to first asthma exacerbation, should be modeled using Cox proportional hazards regression.

Although the outcome variable may determine the type of regression model to use, it may not always be the result of interest from the regression. In some cases, such as clinical prediction models, predicting the outcome is the goal of the model. Often in epidemiology, however, the goal of a regression analysis is to explain the relationship between variables; that is, to quantify the effect between exposure and the outcome, often including covariates to control for confounding. Often, moreover, the implicit goal in these cases is causal inference, and in the right conditions, regression can allow for estimation of causal effects. When explanation is the goal of a model, the regression coefficients, rather than the outcome, are the result of interest. The coefficient for a given covariate represents the magnitude of change in the outcome that can be expected for a 1-unit change in the variable. For linear regression, the regression coefficients can be interpreted directly as the magnitude of change from the covariate; for logistic, Poisson, and Cox regressions, the exponentiated form of the coefficient corresponds with specific measures of effect (Table 4). In univariable models, the coefficient (or exponentiated coefficient) represents the unadjusted (crude) measure of effect; in a multivariable model, coefficients provide the measure of effect for the given variable, adjusted for all other covariates in the model.

Despite their ubiquity in the medical literature, regression models have limitations. Adding too many predictors to the model can lead to separation or overfitting. Separation is a problem in the model fitting algorithm that results in inflated estimates and extremely (sometimes infinitely) wide confidence intervals (Mansournia 2018). Overfitting occurs when the regression model learns to predict the study data so well that it fails to generalize when validated in other data sets. To avoid overfitting, it is generally suggested to include no more than one confounder for every 10 outcome events (the "events-per-variable ratio"), though this useful heuristic may not apply to all models in all circumstances (van Smeden 2016). Therefore, researchers must choose which variables to include in the model; this is an area of ongoing controversy, where pitfalls are common. We discuss it further in later sections.

Beyond these limitations, a chosen regression model should be appropriate for the data being modeled (see Table 4), and each model makes assumptions that sometimes do not hold. There are "model diagnostics" to evaluate whether assumptions have been violated. A discussion of these diagnostics is beyond the scope of this chapter, but ideally, the paper will report how the assumptions were checked; pharmacists should consider the presence or absence of

Table 4. Regression Models and Their Coefficients

Regression Model	Exponentiated Coefficient	Interpretation
Cox	Hazard ratio	The hazard of the outcome increased/decreased by *x* times
Linear	N/A	The outcome increased/decreased by *x* units
Logistic	Odds ratio	The odds of the outcome increased/decreased by *x* times
Poisson	Rate ratio	The rate of the outcome increased/decreased by *x* times

N/A = not applicable.

Information from: Grant SW, Hickey GL, Head SJ. Statistical primer: multivariable regression considerations and pitfalls. Eur J Cardiothorac Surg 2019;55:179-85; Vetter TR, Schober P. Regression: the apple does not fall far from the tree. Anesth Analg 2018;127:277-83.

this information when evaluating a paper reporting regression modeling. One important assumption with Cox models that should be checked is the proportional hazards assumption. Cox models assume proportional hazards between groups (hence the name), and researchers using these models should report how this assumption was evaluated. If the proportional hazards assumption is violated, alternative approaches are necessary for analysis of time-to-event data. Finally, regardless of the model chosen, no regression can account for unmeasured confounders.

Propensity Score Methods

Propensity score methods have gained popularity in recent years, particularly in comparative effectiveness studies. Unlike previously discussed tools for addressing confounding, propensity scores are appealing because they generally overcome the "events-per-variable" problem of multivariable modeling, allowing the inclusion of many more potential confounders. This is a particularly important advantage in studies with low event rates, small sample sizes, or several confounders, where there will be too few events to permit the inclusion of many variables in a regression model.

The propensity score, described as a balancing score, is used to balance pretreatment variables between exposed (treated) and unexposed (control) patients in an observational study (Austin 2011). Specifically, the propensity score is the probability that a subject receives the exposure of interest, conditional on baseline covariates, and ranges from 0 to 1. If all confounding pretreatment variables can be identified and measured, any propensity score-adjusted analysis should result in findings identical to what would have been observed if treatment had been randomized. In addition, because the propensity score can be calculated before data are otherwise analyzed, it separates design from analysis. Because of these features, propensity scores have been described as "the observational study equivalent of complete (i.e., unrestricted) randomization in a randomized experiment"; however, this characterization is only fair when strong assumptions are met (Rubin 2007). Specifically, propensity score analysis assumes that all confounders have been measured and included and that all patients could theoretically have received the treatment. These assumptions should sound familiar because they are the causal identification criteria of conditional exchangeability and positivity. Because of the widespread use of propensity score methods, it is essential that pharmacists understand their applications, strengths, and limitations.

Controlling for confounding using propensity scores is a two-step process. In the first step, the propensity score itself is calculated. The propensity score is commonly estimated using logistic regression, though other methods are sometimes used. The propensity score can thus reduce all included covariates into a single number, which is used to adjust for confounders in the second step of the process. In the second step, the propensity score–adjusted association between exposure and outcome is estimated using one of four approaches: propensity score matching, propensity score stratification, inverse probability of treatment weighting (IPTW), or covariate adjustment. Newer methods for applying propensity scores, such as overlap weights, have also been developed, but we do not generally consider them in this chapter.

In propensity score matching, the propensity score itself is used as the matching variable. Each treated patient is matched to one or more control patients using any of several matching methods. In nearest neighbor matching, the pairs with the closest propensity scores are matched (Austin 2011). With optimal matching, the goal is to minimize the average distance between pairs for the entire sample. Caliper matching uses a prespecified range (the caliper width) and accepts any matches where the distance between a pair's propensity scores is within that range. Simulation data indicate that caliper matching is likely the best method for propensity score methods, using a caliper width of 0.2 (technical meaning: 0.2 standard deviations of the logit of the propensity score). In all cases, unmatched patients are excluded from analysis. Once

the final matched cohort is available, the data are further analyzed to obtain absolute and/or relative measures of effect such as risk differences and risk ratios; however, the paired nature of the data must be considered.

Propensity score stratification is a less common propensity score method used in the literature. Propensity score stratification involves grouping all patients in the data set according to their propensity scores. The recommended number of strata is 5 (i.e., quintiles of the propensity score), though increasing the number of strata can further reduce confounding (Austin 2011). Within each strata, measures of association between exposure and outcome are estimated and then pooled together to generate a single overall estimate. This analysis is the same as if stratifying on individual variables, but using the propensity score allows the inclusion of more confounders than would be possible if stratifying on each combination of variables separately.

In addition to matching or stratifying on the propensity score, researchers may choose to conduct a weighted analysis using the score (Austin 2011). Weighting is a common analytical tool and can be seen, for example, in weighted survey designs, where each response is weighted so that the overall results are representative of the overall population. Weighting by the propensity score, IPTW creates a "pseudopopulation" where the distribution of measured pretreatment covariates is independent of exposure, thereby minimizing confounding from those factors. The weighting technique is one of the least transparent of the propensity score methods. In essence, each subject's weight is the inverse of the propensity score (i.e., the probability that they would have received the treatment). Patients who received treatment despite a low probability of doing so (i.e., a low propensity score) have larger weights than patients who received treatment who were highly likely to be treated (high propensity scores). In untreated patients, the opposite is true: those who were untreated despite a high propensity score have larger weights than those with low propensity scores. These weights are then applied in subsequent analyses to provide (theoretically) unconfounded estimates of treatment effect.

Covariate adjustment involves including the propensity score as a variable in a second statistical model, usually a multivariable regression model. This second multivariable model includes the outcome as the dependent variable and the treatment and propensity score as independent variables. The propensity score theoretically simultaneously adjusts for all the confounders that were included in its estimation. This greatly reduces the number of variables in the second statistical model (dimension reduction), thus reducing the risk of overfitting; however, the second model may still include other variables that were not included in the propensity score. The output of this model will be adjusted measures of effect (e.g., adjusted odds ratios, adjusted hazard ratios). Covariate adjustment will not exclude any patients from the analysis unless additional covariates with missing data are added to the model. This propensity score method is fairly easy to implement, and results are presented and interpreted in a manner similar to other methods in observational studies. However, reliance on a multivariable model means that this approach has all the same challenges of "normal" multivariable regression. Essentially, this approach can fail to adequately adjust for confounding if either the propensity score itself or the second model is incorrect.

Each of the discussed propensity score methods has advantages and disadvantages, some of which have been referenced already, and are summarized in Table 5. In most scenarios, either matching or weighting will be preferred. With limited exceptions in favor of matching, these approaches perform equally well at reducing confounding and outperform stratification or covariate adjustment. Both matching and weighting also have the advantage of allowing for the calculation of both absolute and relative measures of effect in the matched or weighted sample. Results can be reported

Table 5. Advantages and Disadvantages of Propensity Score Methods

Feature	Matching	Stratification	Weighting	Adjustment
Reduces confounding	Yes	Somewhat	Yes	Somewhat
Uses all data	No	Yes	Yes	Yes
Absolute and relative measures of effect	Yes	No	Yes	No
Adjusted measures of effect	Yes	Yes	Yes	Yes
Second statistical model required	No	No	No	Yes
"Black box" analysis[a]	No	No	Yes	Yes

[a]These methods produce interpretable results, but the underlying mathematical model may be difficult to explain and difficult for experts in practical domains to understand.

Information from: Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 2011;46:399-424.

in a manner similar to those of RCTs. Covariate adjustment and stratification only allow for reporting of adjusted relative measures of effect. Between matching and weighting, the choice is likely a matter of data availability and investigator preference. Inverse probability of treatment weighting may be methodologically more of a "black box," but it preserves the entire sample size and does not require accounting for matched data in subsequent analysis.

Instrumental Variables

The last analytical method for addressing confounding and providing conditional exchangeability is instrumental variables. An instrumental variable causes variation in treatment assignment similar to what would be seen from randomization; sometimes, these studies are called "natural randomization" (Maciejewski 2019). The benefit of instrumental variable analysis is that a variable that produces the same effects as randomization should balance both measured and unmeasured confounding. The limited use of these methods in the medical literature may be because of the difficulty finding variables that can plausibly meet the criteria needed to define an instrumental variable. For a variable to be an instrumental variable, it must meet three criteria: (1) it determines a patient's treatment assignment, (2) there is no confounding between the instrumental variable and the outcome, and (3) it should not be associated with the outcome through any pathways other than its effect on treatment assignment (Hernán 2020; Westreich 2019a). The most common way to apply instrumental variables to the adjustment of confounding is through use of two-stage least-squares regression (Ertefaie 2017).

In observational studies, instrumental variables are not always as readily identifiable. Variables commonly used as instruments generally fall in the categories of geographic distance, regional variation, facility variation, physician variation, and calendar time (Maciejewski 2019; Ertefaie 2017). To illustrate a facility variation instrumental variable, consider a study comparing choice of agent used for stress ulcer prophylaxis and risk of pneumonia (Bateman 2013). Some hospitals preferentially use proton pump inhibitors, whereas others use histamine-2 receptor antagonists for stress ulcer prophylaxis. Thus, the hospital the patient was admitted to served as the instrumental variable. The authors assumed that the hospital of admission determined which class of acid-suppressive therapy was received (criterion 1); patients were likely unaware of the hospital's preferred prophylactic agent, and thus there was no confounding between hospital of admission and subsequent pneumonia (criterion 2); and the hospital of admission was not otherwise associated with patients developing pneumonia (criterion 3). The validity of this study's results will depend on the validity of the assumptions regarding the instrumental variable criteria.

If a variable does not meet the criteria discussed earlier, the resulting estimated measures of effect from an instrumental variable analysis will be biased. However, it is rarely possible to prove a variable meets these criteria. Hence, studies using these instrumental variable methods must make assumptions regarding the plausibility of the criteria. Consequently, although instrumental variables may theoretically be able to estimate treatment effects even when unmeasured confounding is present, they do so by relying on unverifiable assumptions (Hernán 2006). One common issue is violation of the second criterion due to the presence of a confounder between the instrumental variable and the outcome, an "instrument-outcome confounder" (Garabedian 2014).

The most significant limitations of instrumental variable methods are therefore the difficulty in finding a plausible instrument and the strong assumptions required for the analysis to be considered unbiased. If a true instrumental variable is identified, it may indeed be "an epidemiologist's dream" (Hernán 2006). When reviewing a study using instrumental variables, a pharmacist should focus on whether the validity of the assumption that the instrumental variable meets the necessary criteria (Ertefaie 2017). Causal diagrams can help the reviewer visualize the potential for violations of these assumptions, such as the presence of potential instrument-outcome confounders (Hernán 2006). The author's justification for choice of instrument, including how it predicts treatment and how it is not associated with the outcome, should also be examined (Maciejewski 2019).

CHOOSING THE RIGHT TOOL(S) FOR THE QUESTION: AVOIDING PITFALLS IN DESIGN AND ANALYSIS

So far, we have covered concepts that are fundamental to contemporary epidemiology and reviewed common tools in study design and analysis that can be used to improve study validity. With the right conditions and certain assumptions, estimates from well-designed studies may allow epidemiologists to move from discussions of association to causation. However, generating these estimates requires the right tools to be deployed in the right context. Being able to recognize common issues and potential pitfalls associated with specific design or analysis choices can help prepare pharmacists to identify potential sources of error in a study.

Pitfalls in the Analysis of Matched Data

Matching is often considered a tool for controlling confounding. Although it has some limitations, matching can make sampling convenient, improve statistical precision, and sometimes help control for unmeasurable variables, such as controlling for genetics by matching siblings or twins (Pearce 2016). However, matching can also result in more serious biases if not analyzed appropriately. The analysis of matched data is complex because the appropriate methods depend on the study design and relationship between variables. However, some general guidance can enable clinicians to identify critical risks in a matched analysis.

In case-control studies, matching generally does not control confounding and should only be done to improve statistical efficiency (Pearce 2016). Matching in case-control studies causes selection bias because the matching variable is often associated with the exposure, and further statistical adjustment will be necessary to address this bias (Mansournia 2013). In addition, if the matching variable was a confounder, it still needs to be adjusted for. The exception is if exposure actually has no effect on disease, in which case no bias will arise. However, it is not as possible to prove an association between exposure and disease is absent; thus, matching variables should generally be adjusted for (Pearce 2016). In matched cohort studies, matching does control for confounding by the matched variables. However, if statistical adjustment of other variables is performed, adjustment for the matched variables is also often necessary (Mansournia 2013). Thus, in general, a reviewer should be suspicious of any matched study that does not also statistically adjust for the matched variables.

Pitfalls in the Analysis of Time-Varying Data

It is easy to think of treatments and outcomes as fixed events: a patient is either treated or not, and the patient either has the outcome or does not. However, both treatments and outcomes can be dynamic events that occur more than once. This creates a problem for most common statistical approaches, which assume that every observation is independent from others. With time-varying data, the same patient could provide data points multiple times, and these data points would be correlated. If the data are regarding the outcome, a repeated-measures analysis is needed (Fitzmaurice 2008). Several different approaches to repeated-measures analysis are available, including change scores, analysis of covariance, generalized linear mixed models, and generalized estimating equations. Although these methods vary in complexity, all can account for the correlated nature of the outcome.

If the repeated data pertain to exposure, time-varying exposure methods are needed. One of the simplest ways to do this is to add exposure as a time-varying covariate to the outcome model. This is one of the recommended strategies for addressing immortal time bias and is commonly done with Cox proportional hazards models. With time-varying covariates, each individual can change from treated to untreated and back as many times as the investigators decide to allow in the model. More complex methods, collectively called "g-methods," can account for both time-varying exposure and confounders (Hernán 2020). The details of these particular methods are beyond the scope of this chapter; for a pharmacist reviewing a study, recognizing that the outcome or exposure has a time-varying element is sufficient to identify whether suboptimal methods have been used.

Pitfalls in the Use of Regression Models

Variable selection in regression models is the subject of ongoing debate and is by far one of the most important pitfalls to be aware of when evaluating a study. To avoid overfitting and separation, building a regression model often involves a process of choosing which variables should be included. The challenge for a researcher or reader is to first understand the goal of the model. If the goal is prediction, any predictor that improves predictive accuracy should be included. In contrast, if the goal is explaining the relationship between variables (causal inference), variable selection needs to be more intentional, given prior knowledge of the topic. Data-driven (i.e., statistical) procedures such as the stepwise approach or change-in-estimate method are widely used for variable selection. In the stepwise approach, variables are added to and/or removed from the model in an iterative process on the basis of a statistical threshold, often a p value threshold, until only "significant" variables remain. With the change-in-estimate approach, covariates are only included in the model if their inclusion changes the estimated effect of the exposure by a prespecified threshold, usually 10%. Although both of these approaches can reduce the number of variables in a model, neither guarantees that only confounders will be included because both mediators and colliders can have "significant" effects in the model. By including mediators or colliders, data-driven variable selection approaches can result in overadjustment and/or selection bias. In addition, both approaches can lead to the omission of important confounders if they do not meet the statistical threshold for inclusion. Sometimes, these omitted variables are put back in the model ("forced in") if they are considered important.

Overall, although data-driven approaches may suffice for prediction, they are inadequate when the goal is causal inference. Newer tools such as shrinkage, Bayesian approaches, and machine learning methods may improve variable selection in the future, though they may still potentially induce bias through inclusion of non-confounding variables. For now, many epidemiologists, particularly those interested in causal inference, advocate abandoning statistical approaches to variable selection in favor of choosing variables according to subject matter expertise. Using causal diagrams to map the relationships among variables can help identify the optimal set of variables needed to adequately control for confounding and estimate causal effects (see Figure 4).

Pitfalls in Propensity Score Analysis

Propensity score methods shift the variable selection problem from modeling the outcome to modeling the treatment. The ability of the propensity score to adequately adjust for confounding depends on accurately predicting the probability of treatment; thus, the challenge of variable selection is of primary importance when estimating the propensity score. Fortunately, propensity score estimation variable selection is perhaps less fraught compared with regression. With

propensity score estimation, the goal is prediction. Therefore, any variable that improves prediction should be evaluated for inclusion in the model used to estimate the score, though some have suggested causal inference principles of variable selection should still apply (Hernán 2020). Regardless of the variable selection approach, three pitfalls must be kept in mind. First, only pretreatment variables should be included in propensity score estimation because only pretreatment variables can influence treatment selection. Second, variables that predict treatment, but are not associated with the outcome (instrumental variables), counterintuitively generate bias in the propensity score; thus, propensity scores should only include confounders and variables associated with the outcome (Brookhart 2006). Finally, from an overfitting perspective, any number of variables meeting these conditions can be included when estimating the propensity score, though mathematical and computational limitations may impose a maximum number of variables according to choice of propensity score estimator and available data.

Pitfalls also exist in applying and interpreting propensity scores. Ultimately, the propensity score should balance treatment groups; hence, "balance diagnostics" should be reported to confirm that balance has been achieved. The balance measure of choice is the absolute standardized difference, which has been shown to perform better than other methods of evaluating covariate balance (Ali 2014). The absolute standardized difference provides an estimate in the differences between mean (for continuous variables) and prevalence/incidence (for categorical variables) before and after propensity score adjustments and can be calculated in studies that use propensity score matching, stratification, or weighting. Alternative diagnostics for covariate adjustment have also been described (Austin 2008). Statistical significance testing of adjusted covariates as the only balance diagnostic in propensity score-matched designs should be discouraged because statistical significance will be influenced by sample size in addition to balance (Austin 2011). Unfortunately, reporting of propensity score methods, including the variables used in the propensity score, method of propensity score adjustment, and balance diagnostics, is inconsistent (Yao 2017; Ali 2015). For a pharmacist to fully evaluate a study using propensity scores, such information must be made available. Without it, it is impossible to assess the appropriateness of methods, risk of bias, and potential for residual confounding. The growing popularity of propensity scores and variability in reporting has led to calls for justification and standardized reporting of methodological details, and a reporting guideline has been proposed (Roth 2019; Yao 2017).

CONCLUSION

The field of epidemiology has significantly advanced over the past century, and although counting exposures and outcomes remains at its core, the methods used have become more sophisticated. Improved understanding of the nature of bias and confounding has allowed for increasingly refined study designs and analytical methods to generate valid estimates of measures of effect. Recent developments in causal inference frameworks have allowed researchers to more clearly specify causal questions, determine conditions that allow for valid estimates of causal effects, and make explicit the assumptions required to make causal inferences from observational data. With intentional design and analysis choices to minimize bias, confounding, and measurement error, assumptions that the causal identification criteria of temporality, consistency, and exchangeability (or at least, conditional exchangeability) may be justified.

The contents of this chapter, which are by no means exhaustive, lay a foundation for pharmacists to advance patient care through a critical appraisal of the literature and evidence-based optimization of therapy. Pharmacists play a key role in understanding and interpreting study results and

Practice Points

As the field of epidemiology continues to advance, methods of conducting valid studies with potential causal implications will continue to improve. Pharmacists evaluating the literature should keep in mind the following key points regarding the current state of epidemiologic thinking:

- **Measures of association continue to be a mainstay in quantifying relationships between exposure and outcome.**
- **All studies are susceptible to bias, including selection bias and information bias. Study design is the best place to address biases. Analysis should take care to avoid adjusting for a collider or mediator because this can cause selection bias and overadjustment, respectively.**
- **Causal identification conditions have been described, and when these are met (or assumed to be met), measures of association become measures of causal effects.**
- **RCTs generally meet these causal identification conditions; thus, results of RCTs can generally be interpreted as causal effects.**
- **For observational studies, confounding remains one of the primary challenges to causal inference. Causal diagrams can help identify which variables should be adjusted for to reduce confounding and which ones should not be adjusted for to avoid inducing bias.**
- **Restriction, matching, and stratification are all relatively simple tools for addressing confounding. However, they can handle only a few confounders and may cause selection bias.**
- **Regression models can control for multiple confounders, but variable selection becomes a problem. If the goal is to understand the effect of exposure on outcome, variable selection should be based on causal diagrams, not data-driven approaches.**
- **Propensity score methods also control for multiple confounders. Propensity score estimation should only include baseline variables and should avoid instrumental variables. Standardized differences should be reported to assess propensity score performance.**

applying them to patient care to improve outcomes. Although the names and terminology used by epidemiologists sometimes differ from what may be used in evidence-based medicine, pharmacists will recognize many of the concepts and tools described in this chapter. These tools are widely used in the medical literature, and with an understanding of how these design and analysis choices can minimize bias and confounding, pharmacists will be well equipped to critically evaluate studies to discern which results are worth incorporating into practice and which are likely the result of systematic errors. Furthermore, by applying causal frameworks to research and practice, we can advance our understanding of which results from nonrandomized studies represent potentially causal effects.

REFERENCES

Ali MS, Groenwold RHH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015;68:122-31.

Ali MS, Groenwold RHH, Pestman WR, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. Pharmacoepidemiol Drug Saf 2014;23:802-11.

Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 2011;46:399-424.

Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf 2008;17:1202-17.

Bateman BT, Bykov K, Choudhry NK, et al. Type of stress ulcer prophylaxis and risk of nosocomial pneumonia in cardiac surgical patients: cohort study. BMJ 2013; 347:f5416.

Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149-56.

Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Health 2004;58:635-41.

Ertefaie A, Small DS, Flory JH, et al. A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2017;26:357-67.

Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020;158:S21-8.

Fitzmaurice GM, Ravichandran C. A primer in longitudinal data analysis. Circulation 2008;118:2005-10.

Garabedian LF, Chu P, Toh S, et al. Potential bias of instrumental variable analyses for observational comparative effectiveness research. Ann Intern Med 2014;161:131.

Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. Anesthesiology 2020;132:951-67.

Gerhard T. Bias: considerations for research practice. Am J Health Syst Pharm 2008;65:2159-68.

Glass TA, Goodman SN, Hernán MA, et al. Causal inference in public health. Annu Rev Public Health 2013;34:61-75.

Grant SW, Hickey GL, Head SJ. Statistical primer: multivariable regression considerations and pitfalls. Eur J Cardiothorac Surg 2019;55:179-85.

Hartung DM, Touchette D. Overview of clinical research design. Am J Health Syst Pharm 2009;66:398-408.

Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. Am J Public Health 2018;108:616-9.

Hernán MA. Invited commentary: selection bias without colliders. Am J Epidemiol 2017;185:1048-50.

Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615-25.

Hernán MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006;17:360-72.

Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295-300.

Holman DJ, Arnold-Reed DE, de Klerk N, et al. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. Epidemiology 2001;12:246-55.

Hulley SB, Cummings SR, Browner WS, et al., eds. Designing Clinical Research, 4th ed. Wolters Kluwer/Lippincott Williams & Wilkins, 2013.

Lash TL, VanderWeele TJ, Haneuse S, et al. Modern Epidemiology, 4th ed. Lippincott Williams & Wilkins, 2021.

Lederer DJ, Bell SC, Branson RD, et al. Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. Ann Am Thorac Soc 2019;16:22-8.

Livorsi DJ, Linn CM, Alexander B, et al. The value of electronically extracted data for auditing outpatient antimicrobial prescribing. Infect Control Hosp Epidemiol 2018;39:64-70.

Maciejewski ML, Brookhart MA. Using instrumental variables to address bias from unobserved confounders. JAMA 2019;321:2124.

Maclure M, Schneeweiss S. Causation of bias: the episcope. Epidemiology 2001;12:114-22.

Magill SS, Edwards JR, Bamberg W, et al. Multistate point-prevalence survey of health care–associated infections. N Engl J Med 2014;370:1198-208.

Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic regression: causes, consequences, and control. Am J Epidemiol 2018;187:864-70.

Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. Int J Epidemiol 2013;42:860-9.

Mansournia MA, Higgins JPT, Sterne JAC, et al. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology 2017;28:54-9.

Navalkele B, Pogue JM, Karino S, et al. Risk of acute kidney injury in patients on concomitant vancomycin and piperacillin-tazobactam compared to those on vancomycin and cefepime. Clin Infect Dis 2017;64:116-23.

Pearce N. Analysis of matched case-control studies. BMJ 2016;352:i969.

Porta M, ed. A Dictionary of Epidemiology, 5th ed. Oxford University Press, 2008.

Roth JA, Juchler F, Widmer AF, et al. Plea for standardized reporting and justification of propensity score methods. Clin Infect Dis 2019;68:710-1.

Rothman KJ. Epidemiology: An Introduction, 2nd ed. Oxford University Press, 2012.

Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26:20-36.

Saver JL, Lewis RJ. Number needed to treat: conveying the likelihood of a therapeutic effect. JAMA 2019;321:798.

Sedgwick P. Explanatory trials versus pragmatic trials. BMJ 2014;349:g6694.

Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008;8:70.

Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;l4898.

Suissa S. Immortal time bias in pharmacoepidemiology. Am J Epidemiol 2008;167:492-9.

Suissa S, Dell'Aniello S. Time-related biases in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2020;29:1101-10.

Suissa S, Dell'Aniello S, Vahey S, et al. Time-window bias in case-control studies: statins and lung cancer. Epidemiology 2011;22:228-31.

van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016;16:163.

Vetter TR, Mascha EJ. Bias, confounding, and interaction: lions and tigers, and bears, oh my! Anesth Analg 2017;125:1042-8.

Westreich D. Epidemiology by Design: A Causal Approach to the Health Sciences. Oxford University Press, 2019a.

Westreich D, Edwards JK, Lesko CR, et al. Target validity and the hierarchy of study designs. Am J Epidemiol 2019b;188:438-43.

Yao XI, Wang X, Speicher PJ, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. JNCI J Natl Cancer Inst 2017;109:djw323.

Self-Assessment Questions

Questions 1–3 pertain to the following case.

T.S., a pharmacy resident, is designing a research project to compare the risk of surgical site infection (SSIs) after colorectal surgery in patients receiving ertapenem or cefoxitin as antimicrobial surgical prophylaxis. T.S. expects to collect data on 200 procedures, with SSIs expected to occur in 7% of cases. Patient age, preoperative albumin, preoperative American Society of Anesthesiologists class (a measure of risk of postoperative complications), Charlson Comorbidity Index, and procedure duration are identified as potential confounders. T.S. wants to use propensity score methods in this study and prefers to report both absolute and relative measures of effect in the adjusted analysis while minimizing any loss in sample size.

1. Which one of the following propensity score methods would be most appropriate for T.S. to use in this analysis?

   A. Matching with 1:1 matching
   B. Stratification
   C. Weighting (inverse probability of treatment weighting [IPTW])
   D. Adjustment (covariate adjustment)

2. T.S.'s use of propensity score methods in this study is intended to ensure that, conditional on the adjusted confounders, which one of the following causal identification criteria is most likely to be met?

   A. Consistency
   B. Exchangeability
   C. Positivity
   D. Temporality

3. When reporting the results of the propensity score analysis, which one of the following will best facilitate readers' ability to assess the validity of the methods and results of T.S.'s analysis?

   A. 95% confidence intervals
   B. P values
   C. Regression coefficients
   D. Standardized differences

Questions 4 and 5 pertain to the following case.

The REMCOV retrospective cohort study compared survival among critically ill patients who received remdesivir with survival among patients who did not. All patients admitted to the ICU with COVID-19 were included in the cohort. Patients were included in the remdesivir arm if they received the drug at any point during admission; otherwise, patients were included in the control group. The end point in REMCOV was time to death, evaluated using a multivariable Cox proportional hazards model. The authors reported an unadjusted hazard ratio of 0.47 (95% CI, 0.30–0.74) and an adjusted hazard ratio of 0.60 (95% CI, 0.40–0.90). Variables adjusted for included age, pretreatment pneumonia severity index, D-dimer, absolute lymphocyte count, and any corticosteroid use.

4. Which one of the following best interprets the REMCOV adjusted hazard ratio result?

   A. Remdesivir is associated with a nonsignificant 40% decrease in hazard of death.
   B. Remdesivir is associated with a significant 40% decrease in hazard of death.
   C. Remdesivir is associated with a nonsignificant 60% increase in hazard of death.
   D. Remdesivir is associated with a significant 60% increase in hazard of death.

5. Given the described methods, which one of the following best evaluates potential sources of bias in the REMCOV study?

   A. Possible selection bias from controlling for pretreatment pneumonia severity index
   B. Possible immortal time bias from how treatment was categorized
   C. Possible information bias from subjective assessment of end point
   D. Bias unlikely on the basis of described study methods

6. CANA1C was a randomized, double-blind, placebo-controlled trial of canagliflozin added to background therapy in type 2 diabetes. The authors found that euglycemic diabetic ketoacidosis occurred in 10 of 10,687 patients randomized to canagliflozin and 2 of 6909 randomized to placebo. The total CANA1C patient follow-up time was 15,526 person-years in the canagliflozin group and 8583 person-years in the placebo arm. Which one of the following most accurately depicts the CANA1C incidence rate ratio for euglycemic diabetic ketoacidosis with canagliflozin relative to placebo?

   A. 0.3
   B. 0.4
   C. 2.8
   D. 3.2

Questions 7–9 pertain to the following case.

K.T. conducts a case-control study evaluating the risk of parasitic infections in patients with asthma treated with any of the new anti–interleukin 5 (anti–IL-5) agents (benralizumab, mepolizumab, or reslizumab). She uses a cumulative sampling strategy: K.T. first identifies patients with asthma who developed a parasitic infection and then selects a corresponding group of patients with asthma who did not develop a parasitic infection at any point during the study. K.T. then

looks back and groups the patients according to whether they were prescribed an anti–IL-5 agent.

7. To control for confounding in her study, K.T. matches cases to controls on the basis of age, oral corticosteroid use, and state of residence. Which one of the following best evaluates how this will affect the validity of K.T.'s results?

   A. Induce immortal time bias
   B. Induce information bias
   C. Induce selection bias
   D. Unlikely to affect validity of results

8. K.T. obtains data from an insurance claims database and identifies parasitic infection using unvalidated ICD-10 diagnosis codes. Which one of the following best evaluates how this will affect the validity of K.T.'s results?

   A. Induce immortal time bias
   B. Induce information bias
   C. Induce selection bias
   D. Unlikely to affect validity of results

9. K.T.'s study reports an odds ratio for parasitic infection of 5.0 (95% CI, 0.45–55.58) with anti–IL-5 therapy compared with no anti–IL-5 therapy. Which one of the following most appropriately interprets this result?

   A. Anti–IL-5 therapy nonsignificantly decreases the odds of parasitic infection by 5 times.
   B. Anti–IL-5 therapy nonsignificantly increases the odds of parasitic infection by 5 times.
   C. Anti–IL-5 therapy significantly increases the odds of parasitic infection by 5 times.
   D. Anti–IL-5 therapy significantly increases the odds of parasitic infection by 5 times.

Questions 10 and 11 pertain to the following case.

ESKLEB was an open-label randomized clinical trial (RCT) in which patients with bacteremia caused by ceftriaxone-nonsusceptible *Escherichia coli* or *Klebsiella* spp. were assigned prospectively to definitive therapy with piperacillin/tazobactam or meropenem. Block randomization was performed and stratified according to the infecting pathogen, infection source, and disease severity. The randomization sequence was provided using an online portal. The ESKLEB primary outcome of 30-day all-cause mortality occurred in 23 of 187 patients in the piperacillin/tazobactam group and 7 of 191 patients in the meropenem group.

10. Which one of the following most accurately depicts the risk ratio for treatment failure with piperacillin/tazobactam compared with meropenem in ESKLEB?

    A. 3.3
    B. 3.6
    C. 3.7
    D. 11.1

11. Which one of the following features of the ESKLEB study is most effective for minimizing confounding of the estimated measure of association in the intention-to-treat analysis?

    A. Allocation concealment
    B. Open-label design
    C. Prospective design
    D. Randomization

Questions 12 and 13 pertain to the following case.

R.K. is a pharmacist conducting a retrospective cohort study comparing 30-day treatment failure (yes or no) in patients with gram-negative bacteremia receiving an oral β-lactam or oral fluoroquinolone as stepdown therapy. A total of 163 patients are included in R.K.'s study, with 68 in the β-lactam group and 95 in the fluoroquinolone group. Treatment fails in eight patients, four in each group. Potential confounders include age, biological sex, state of residence, Pitt bacteremia score, and infection source.

12. In R.K.'s study, which one of the following most accurately depicts the unadjusted odds ratio for treatment failure in the β-lactam group compared with the fluoroquinolone group?

    A. 0.7
    B. 1.4
    C. 1.7
    D. 58

13. Given the limited number of treatment failures, which one of the following methods of controlling for all the identified potential confounders would be most appropriate for R.K. to use?

    A. Matching
    B. Multivariable logistic regression
    C. Propensity score weighting (IPTW)
    D. Stratification

Questions 14 and 15 pertain to the following case.

GRAMBAC was an open-label RCT in which patients with gram-negative bacteremia were randomly assigned to receive 7 days (short course) or 14 days (standard course) of antibiotic therapy. Randomization was concealed using sealed, opaque envelopes that were opened sequentially. The GRAMBAC outcome was treatment failure, defined as a composite of all-cause mortality, relapse or other complication, and hospital readmission, all measured 90 days after randomization. Treatment failure occurred in 105 of 306 patients in the short-course treatment arm and 140 of 298 patients in the standard-course treatment arm, for a reported risk difference of -8.2% (95% CI, -16.3% to -0.1%).

14. Which one of the following features of the GRAMBAC study most increases the possibility of information bias?

    A. Allocation concealment
    B. Open-label design
    C. Prospective design
    D. Randomization

15. Which one of the following best interprets the reported measure of effect in GRAMBAC?

    A. Short-course treatment is associated with a nonsignificant 8.2% decrease in the rate of treatment failure.
    B. Short-course treatment is associated with a nonsignificant 8.2% decrease in the risk of 90-day treatment failure.
    C. Short-course treatment is associated with a significant 8.2% decrease in the absolute risk of 90-day treatment failure.
    D. Short-course treatment is associated with a significant 8.2% decrease in the odds of treatment failure.