Standard Review Article

OPEN

Simple design and analysis strategies for solving problems in observational orthopaedic clinical research

Kelsey E. Brown, MD[a], Michael J. Flores, MD[a], Gerard Slobogean, MD[b], David Shearer, MD, MPH[a], Ida Leah Gitajn, MD, MS[c], Saam Morshed, MD, PhD[a,*]

Summary: Randomized controlled trials are the gold standard to establishing causal relationships in clinical research. However, these studies are expensive and time consuming to conduct. This article aims to provide orthopaedic surgeons and clinical researchers with methodology to optimize inference and minimize bias in observational studies that are often much more feasible to undertake. To mitigate the risk of bias arising from their nonexperimental design, researchers must first understand the ways in which measured covariates can influence treatment, outcomes, and missingness of follow-up data. With knowledge of these relationships, researchers can then build causal diagrams to best understand how to control sources of bias. Some common techniques for controlling for bias include matching, regression, stratification, and propensity score analysis. Selection bias may result from loss to follow-up and missing data. Strategies such as multiple imputation and time-to-event analysis can be useful for handling missingness. For longitudinal data, repeated measures allow observational studies to best summarize the impact of the intervention over time. Clinical researchers familiar with fundamental concepts of causal inference and techniques reviewed in this article will have the power to improve the quality of inferences made from clinical research in orthopaedic trauma surgery.

1. Introduction

The randomized controlled trial (RCT) is the gold standard for demonstrating efficacy of an intervention in clinical research. RCTs are commonly used to study causal relationships because randomization is able to balance known and unknown confounders that lead to bias in other observational study designs.[1] However, RCTs are time consuming and expensive to conduct.[1] In some cases, researchers are unable to conduct RCTs because of lack of resources or clinical equipoise and must therefore rely on observational (nonrandomized) study designs. As encountered with prospective RCTs, researchers may encounter unfamiliar problems such as proper handling of repeated measures or loss to follow-up. This article summarizes the 2021 Orthopaedic Trauma Association Basic Science Focus Forum's Symposium on "Simple Design and Analysis Strategies for Solving Common Problems in Orthopaedic Clinical Research," which provided orthopaedic surgeons and researchers with a primer on clinical research methodology to optimize inferences from observational data.

2. Understanding Causal Relationships

The aim of most clinical research is to estimate causality, or a cause-and-effect relationship, between certain variables and/or outcomes. Even when a study is conducted to identify "modifiable risk factors" or predictors of a particular outcome, some plausible causal relationship is implied. Causality has been studied for more than 3 centuries. The enlightenment philosopher David Hume defined a cause as "an object, followed by another,...where, if the first object had not been, the second had never existed."[2,3] Causal inference is the study and application of strategies that allow researchers to make causal conclusions based on scientific data, rather than associations or distributions alone.[4] Central to the notion of causality in clinical research is the concept of a "counterfactual."[3] Theoretically, understanding the effect of an intervention requires the concept of an unobserved situation in which a study participant would go back in time and be given a treatment/exposure/risk factor that they did not receive, then be followed forward in time for their outcome to be recorded, and ultimately be compared with that which was actually observed. In reality, the closest approximation of such a counterfactual scenario is a control group in a randomized experiment. In observational clinical studies, study participants have unique host, injury, and environmental factors/variables that potentially affect both their likelihood of receiving a treatment and likelihood of outcome of interest, and this leads to a mixing of effects known as confounding bias. Randomized control trials (RCTs) minimize bias through randomizing not just treatment but also the distribution of confounding factors that are both known and unknown, thereby neutralizing their ability to confound estimates of effect. However, RCTs are often impossible in a clinical setting because of resource

The authors declare no conflicts of interest.

[a] Department of Orthopedics, University of California San Francisco, San Francisco, CA, [b] Department of Orthopedics, University of Maryland Medical System, Baltimore, MD; and, [c] Department of Orthopedics, Dartmouth Geisel School of Medicine, Lebanon, NH.

* Corresponding author. Address: Saam Morshed, MD, PhD, Department of Orthopedics, University of California San Francisco, 2550 23rd St, San Francisco, CA 94110. E-mail: saam.morshed@ucsf.edu

No funding was received in relation to this manuscript.

OTAI (2023) e239

Received: 28 September 2022 / Accepted: 14 December 2022
Published online 4 May 2023

http://dx.doi.org/10.1097/OI9.0000000000000239

constraints and ethical concerns. Other causal inference methods can provide means to best simulate the counterfactual environment in the setting of observational clinical studies by providing safeguards in study design and/or analysis against bias. To learn these techniques, we must first review 3 foundational causal relationships.

There are 3 fundamental ways in which variables can interact with one another through (1) confounding, (2) mediation, and (3) collision (Fig. 1). The first and most common is that of confounding, which occurs when a variable influences both the intervention and the outcome. At the most basic level, confounding can be defined as a "mixing of effects of extraneous factors (called confounders) with the effect of interest."[3,5] Confounding occurs when investigators fail to identify or estimate fundamental causal parameters from the study data that determine their observations[6] (Fig. 1A). The principle of confounding may best be conceptualized through illustration. In a recent study by Patterson et al,[7] early reoperation rates after internal fixation were compared after either open or closed reduction in 234 patients with femoral neck fractures. In this instance, the reduction method was the intervention (A) (see Fig. 1A) and early reoperation was the outcome (B). Age is just one of many factors that may have influenced choice of reduction technique and also affected outcome; therefore, age can be considered a confounder (C) of the association between reduction technique and reoperation.

Mediation is the situation in which an outside variable lies along the causal pathway of the study variables (Fig. 1B). Blocking a mediator results in the loss of association between the intervention and outcome. This concept is often misunderstood in clinical research and leads to erroneous adjustment. For example, some have been concerned that the study by Patterson et al did not control for quality of reduction.[8] However, open reduction is more likely to result in a better quality of reduction than closed reduction, and a better quality of reduction mediates a better outcome. Controlling for reduction quality would mask the potential effect of the treatment on the outcome by negating the associating of reduction quality with the reduction method and outcome.

Collider bias results from conditional sampling bias or adjustment for a factor that is influenced by 2 independent variables and is a more complex concept (Fig. 1C). Although it will not be discussed in detail here, the most common example of collider bias is Berkson paradox.[9] Adjustment for a collider will induce a spurious association between the 2 independent variables that affect the collider. In summary, estimating causal effects from observational data requires recognition of potential confounding bias as well as causal mediators and potential colliders. Although adjustment for confounders is critical, adjustment for mediators and colliders can introduce new sources of bias that are difficult to predict.

3. Organizing Causal Diagrams

One key tool for identifying and controlling sources of bias is the creation of a causal diagram which uses knowledge of subject matter and of the fundamental relationship described above. Casual diagrams map causal relationships between the intervention(s), outcome(s), and any other related variables conceptualized by the study team. These diagrams provide a clear map of potential confounding variables that should be controlled for and mediating variables within a causal pathway that should not be adjusted for. One effective method of using a causal diagram is to look for "backdoor paths" or constituted by potential confounders that affect both the treatment and outcome. This method involves identifying all the potential variables that affect the intervention and control for those variables if they also have a causal path to the outcome. It is essential to remember that only 1 variable along any backdoor path needs to be controlled for to mitigate the path's bias on the study outcome. This conserves statistical power while still minimizing study bias.

An example of a causal diagram is presented in Fig. 2A. This causal diagram is based on a study by Slobogean et al, where early pain and functional outcomes were compared in minimally displaced complete lateral compression pelvic fractures undergoing operative fixation versus nonoperative management. In this study, many factors including age, injury severity, fracture displacement, and associated injuries were identified as variables affecting the study intervention and/or outcome. Fig. 2B shows a possible scenario where sufficient confounding control is achieved by backdoor pathways adjustment (age, injury severity, and fracture displacement). Another possible scenario is presented in Fig. 2C. In this case, if injury severity is not known or adequately measured, associated injuries can also be controlled for to block the same backdoor pathway as injury severity, to the outcome.

4. Adjusting for Confounding Bias

Once confounding variables have been identified on a causal diagram, there are multiple methods that can be used to control for them. One can either prevent confounding in study design or control for confounding in analysis. As mentioned above, randomization is the gold standard design method for dealing with confounding, as RCTs create the most counterfactual-like environment by randomizing both known and unknown confounders. When a RCT is not appropriate, restriction and matching are effective tools to mitigate bias in observational studies. Restriction involves limiting a study sample to participants with only one level of a potentially confounding factor, thereby eliminating variance in the variable as a potential source of confounding bias. One example of restriction is creating an age cutoff. Patterson et al did not include older patients (older than 65 years) to at least partially mitigate confounding by age. One downside to restriction is that it limits sample size and limits the scope of conclusions to only those patients included in this study. Matching involves choosing control patients in a study based on similar potentially confounding characteristics. This method has many potential limitations, including decreasing the amount of patients that can be included in a study and the number of variables that can be analyzed, leading to decreased statistical power.

Confounding can also be adjusted for in the analysis phase of a study. Two examples of adjustment in the analysis phase include regression and stratification. Stratification involves dividing study participants into groups for analysis. This method ensures that an analysis is performed with only participants with similar characteristics defining the subgroup. Averaging effects over subgroups of the variable of interest may be performed if those effects are sufficiently homogeneous, to yield average causal effects that are no longer biased by the subgrouping variable. The limitation of stratification arises when considering multiple factors that one would like to adjust for simultaneously. Regression is a method used to model the causal relationship between a dependent variable (eg, study outcome) and the independent variable(s) (eg, intervention and confounding variables). The most common regression technique is linear regression. Linear regression involves fitting a linear equation to the data. When adequately fit to the data, regression can provide

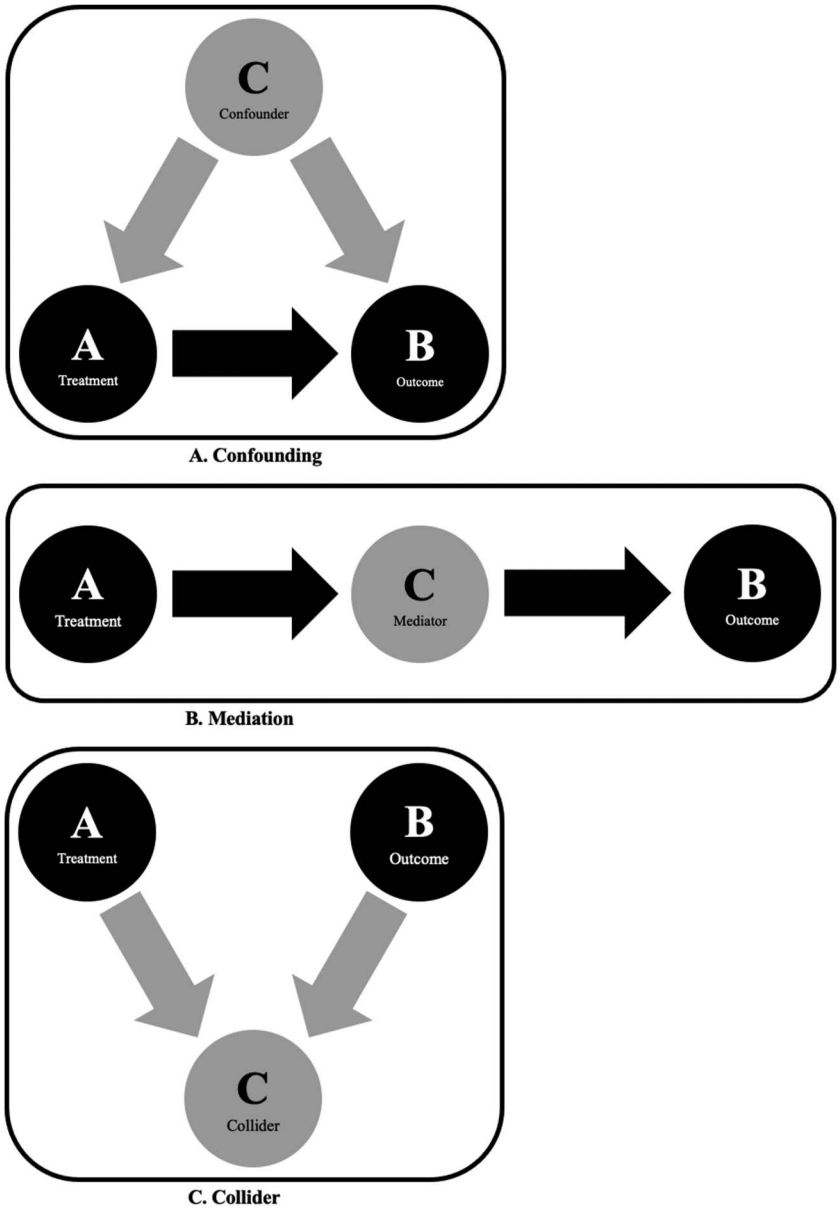
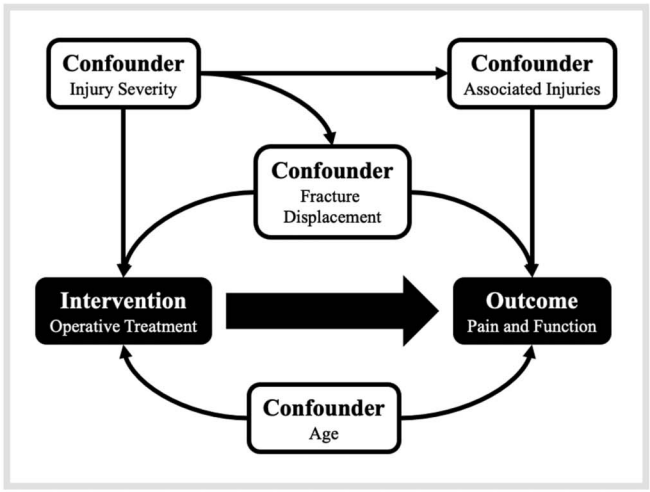


Figure 1. Causal diagrams for fundamental causal relationships between a treatment (A), outcome (B), and a third variable (C). a—The causal relationship between A and B is confounded by variable C. b—The causal relationship between A and B is mediated by variable C. c—No causal relationship exists between A and B, unless one is induced by conditioning on variable C, a collider.

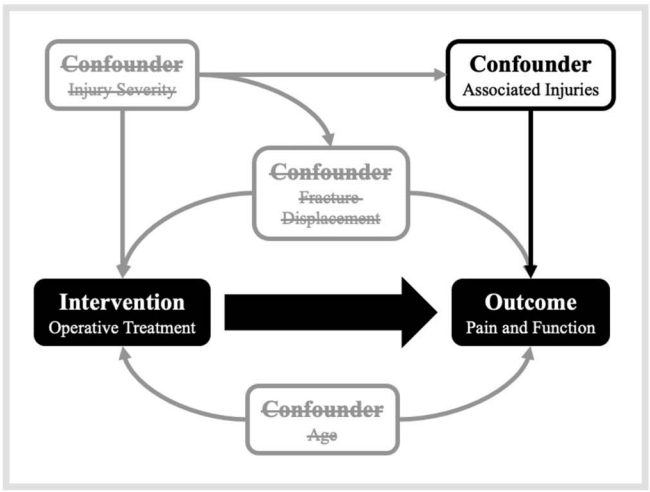
the average causal effect attributable to the intervention of interest holding other factors in the model constant, with much greater efficiency than multidimensional stratification.

Another class of analytical approaches to adjustment for confounding requires an understanding of the drivers of treatment allocation. The *treatment mechanism* is defined as the various factors that influence the decision to provide a specific intervention. For example, in the case of open versus closed reduction for treatment of femoral neck fractures, many variables can influence the decision of the reduction method including patient, surgeon, and injury factors. One key difference between observational data and a RCT is that the treatment method is random in a RCT, whereas the treatment method is not known in an observational study. However, with careful identification of the known and potential confounders influencing treatment with the use of a causal diagram, the treatment mechanism can be estimated with statistical models such as logistic regression. A propensity score (PS) represents the probability between 0 and 1 (0%–100%) that an individual patient would get the intervention of interest, as calculated by a treatment regression model. The PS considers all the factors affecting treatment instead of adjusting for multiple confounding variables one at a time.

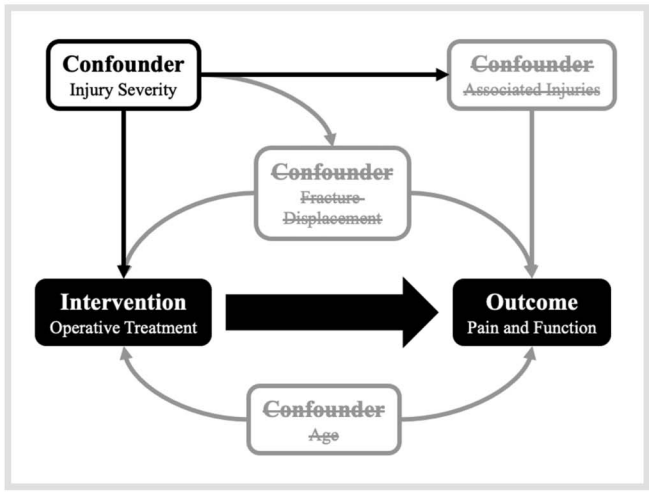
PSs have many quantitative and qualitative uses. PSs between participants can be compared and used for adjustment methods such as matching, stratification, and regression. Adjustment through the use of PSs simulates a counterfactual environment afforded by randomization whereby subjects are equally likely to receive treatment within strata of the PS. In addition, PS distributions can be compared between intervention groups to identify outliers or subgroups of subjects where treatment



A. Causal Diagram Example



B. Control Scenario 1



C. Control Scenario 2

Figure 2. Examples of how causal diagrams can be used to guide efficient "back-door" adjustment for confounding bias of the relationship between an operative treatment and pain and functional outcomes. A—The complete causal diagram with bolded arrow signifying the effect of interest as well as extraneous factors that confound this association through "back-door" pathways. Sufficient confounding control can be achieved by controlling for the minimum subset of confounders that block all "back-door" paths between intervention and outcome. This can be performed validly in this case by adjusting for age, fracture displacement, and either injury severity (B—scenario 1) or associated injuries (C—scenario 2).

assignment may be deterministically assigned and comparison is not appropriate (ie, PS very close to either 0 or 1). Patterson et al used a logistic regression model to estimate PSs for each patient who were used to control for confounding bias in reporting a 2.4-fold increase in the propensity-adjusted hazard of reoperation in the patients undergoing ORIF (95% confidence interval, 1.3–4.4; $P = 0.004$).

5. Missing Data

Missingness and missing data are a normal part of all medical research. Missing data are described as "values that are not available and that would be meaningful for analysis if they were observed."[10] Missed visits, uncollected unique clinical variables, loss to follow-up, participant refusal, and study design issues are all reasons that data could be missing from a study. Missingness can result in reduced sample size, which leads to reduced statistical power, challenges with subgroup analysis and difficulty drawing sound conclusions. It can also lead to various types of bias. It is critical that study teams have an a priori analytic plan in place before starting the study to avoid model dependence and anticipate missingness.

To address missingness, we must first define the types of missingness. Missingness Completely at Random (MCAR) is the probability that the variable missing is unrelated to the value of that variable or other variables in the data set. An example of MCAR would be participants missing patient satisfaction values with no systematic pattern. Missing at Random (MAR) is the probability that the variable missing is unrelated to the value of that variable after controlling for other variables in the data set. For example, MAR would be patient satisfaction data missing mostly among patients with public insurance. Missing Not At Random (MNAR) is the probability that missing data are related to the value of that variable. If only high scoring patient satisfaction data were missing, that would be an example of MNAR. The first 2 types of missingness (MCAR and MAR) can be statistically tested for, but it is difficult to test for and identify MNAR.

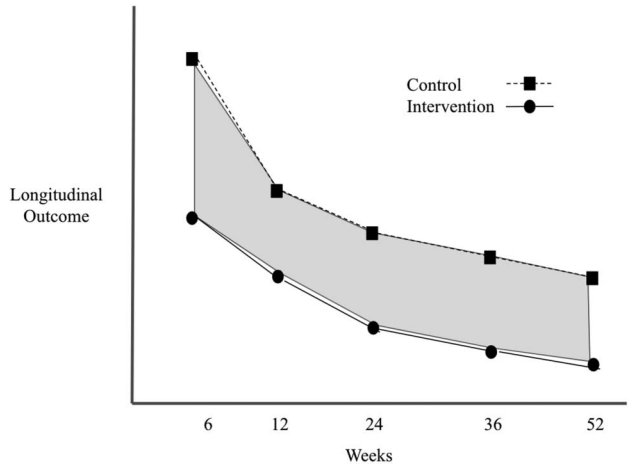


Figure 3. Repeated measurements of longitudinally measured outcome allows for an averaging of difference in scores over multiple time points (shaded area). This averaging of measurements over time increases the statistical power to detect between group differences among a given sample.

The simplest way to address missingness is through a complete record analysis, which restricts the analysis to record with complete data. Although it is the simplest, it often results in a loss of power and, potentially, selection bias. This analysis also requires authors to prove that only cases with complete data are relevant and that you have MCAR. Another less common method to deal with missingness is through single imputation. This method inputs missing values based on other values of the missing variable. For example, the last or baseline observation could be carried forward or analysts could do a mean imputation to fill in the missing value. This method can also introduce bias and is not usually recommended.

Multiple imputation is a very common method to address MCAR and MAR. This method replaces missing values with multiple values from an approximate distribution for missing values that is based on observed data. Essentially, the statistical software of choice creates 50–100 data sets with different values for missing data based on existing data and some degree of randomness. The goal of this method was not to estimate missing values but rather produce unbiased estimates for the population parameters of interest while incorporating a certain degree of uncertainty introduced by a missing variable. This method of addressing missingness was used in a study looking at intramedullary nailing versus external fixation in the treatment of open tibial fractures in Tanzania.[11] This study used multiple imputation as a sensitivity analysis to check for differences that may have arisen due to exclusion of participants with missing observations of interval quality of life data, when estimating main effects with complete case data. In this case, no difference in estimates between the complete case and multiple imputation analyses supported the robustness of their results to those missing data.

Time-to-event analysis is helpful to address missingness because of loss to follow-up also known as censoring. This analysis looks at whether the outcome occurs and when the event occurred. Time-to-event analysis allows investigators to analyze incomplete data by setting the time-to-event as the primary outcome rather than the presence or absence of the event. Every participant then contributes person-time to the analysis as long as they are followed and free of the event of interest (eg, failure or reoperations). A Kaplan–Meier survival curve can be used to graphically demonstrate event-free survival, and comparisons can be made between different treatment groups. Cox proportional hazard models are multivariate time-to-event analyses that allow assessment of treatment effects while controlling for potentially confounding factors. In their article on risk of reoperation after internal fixation of displaced femoral neck fractures, Patterson et al[7] used a Cox proportional hazard model to estimate the hazard of reoperation. This allowed the authors to incorporate data from patients who were lost to follow-up, thereby mitigating selection bias because of censoring.

6. Longitudinal Outcomes Data

Longitudinal outcomes or repeated measures data, such as pain, health-related quality of life measures, and functional outcomes taken at multiple time points for each participant,[12] can be dealt with in a multitude of ways. As these outcomes are usually continuous, 1 option would be to conduct a hypothesis test such as the *t* test at each time point to determine the difference in mean scores at more than 1 time point. However, multiple *t* tests would require additional statistical testing to mitigate the risk of false discovery or Type 1 error. Moreover, *t* tests assume that the observations are independent of one another, which in the case of longitudinal outcomes data is not true. For longitudinal outcomes data, repeated measures analysis would be a better statistical analysis plan.

Repeated measures or multiple observations of an outcome per patient over time form a line of each patient's outcome data. Each patient's line can then be averaged within treatment groups to get the overall treatment effect of each group (Fig. 3). A few examples of techniques that are appropriate for analyzing longitudinal data include mixed effect regression models and generalized estimating equations. One benefit of using repeated measure analysis is that it increases statistical power because an increased number of measurements per patient reduces variance and increases precision of estimates. Repeated measures also allow for inclusion of patients with partially missing data because data from prior and later visits can be used to help estimate data from missing time points. This statistical method also avoids multiple testing and allows analysts to easily account for confounding variables.

The Canadian Orthopaedic Trauma Society (COTS) Clavicle RCT is 1 frequently cited example of how repeated measures analysis should have been used. This trial compared patient outcomes and complication rates of displaced midshaft clavicle fractures after nonoperative treatment and plate fixation.[13] Their primary outcome was the Disabilities of the Arm, Shoulder, and Hand (DASH) score, which was collected from patients at the 6-, 12-, 24-, and 52-week time points.[13] Although this study used multiple *t* tests to show that operative management was superior over time, repeated measures would have been a more valid statistical tool. Repeated measures would have allowed the COTS group to report that, on average, operative treatment improved DASH scores over the first year after surgery. Although the outcome could be reported as an averaged over time, the averaged regression can also be used to estimate point estimates of effect at isolated time points as well while accounting for possible Type 1 error.

7. Conclusion

In this overview article, we have summarized some of the challenges encountered in clinical research as they relate to mitigation of bias. Causal diagrams have been introduced to better conceptualize and differentiate confounders from mediators and colliders in observational data. Once sources of confounding bias have been identified, then steps can be

taken in either the design or analysis phase of this study to adjust for them in estimating average causal effects. Furthermore, multiple imputation and time-to-event analysis can help researchers address missingness and loss to follow-up. Finally, we have highlighted some of the advantages of using longitudinal outcome measurement and discussed how they may be most appropriately analyzed with repeated measure analysis. While not unique to observational studies, understanding the problems and their solution highlighted in this review will improve the quality of this most common clinical study design in orthopaedic trauma surgery.

References

1. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. *BJOG Int J Obstet Gynaecol*. 2018;125:1716.
2. Hume D. *An Enquiry Concerning the Human Understanding: And an Enquiry Concerning the Principles of Morals*. Oxford, England: Clarendon Press; 1894.
3. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14:29–46.
4. Pearl J An introduction to causal inference. *Int J Biostat*. 2010;6:7.
5. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359:248–252.
6. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15:413–419.
7. Patterson JT, Ishii K, Tornetta P, et al. Open reduction is associated with greater hazard of early reoperation after internal fixation of displaced femoral neck fractures in adults 18-65 years. *J Orthop Trauma*. 2020;34:294–301.
8. Wu SH, Mei J. Letter to the editor on "open reduction is associated with a greater hazard of early reoperation after internal fixation of displaced femoral neck fractures in adults 18-65 years". *J Orthop Trauma*. 2020;34:e434.
9. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiol Camb Mass*. 2012;23:159–164.
10. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012;367:1355–1360.
11. Haonga BT, Liu M, Albright P, et al. Intramedullary nailing versus external fixation in the treatment of open tibial fractures in Tanzania: results of a randomized clinical trial. *J Bone Joint Surg Am*. 2020;102:896–905.
12. Sullivan LM. Repeated measures. *Circulation*. 2008;117:1238–1243.
13. Canadian Orthopaedic Trauma Society. Nonoperative treatment compared with plate fixation of displaced midshaft clavicular fractures. A multicenter, randomized clinical trial. *J Bone Joint Surg Am*. 2007;89:1–10.