Chapter 7. Covariate Selection

Brian Sauer, Ph.D.
University of Utah School of Medicine, Salt Lake City, UT

M. Alan Brookhart, Ph.D.
University of North Carolina at Chapel Hill
Gillings School of Global Public Health
Chapel Hill, NC

Jason A. Roy, Ph.D.
University of Pennsylvania, Philadelphia, PA

Tyler J. VanderWeele, Ph.D.
Harvard School of Public Health, Boston, MA

Abstract

This chapter addresses strategies for selecting variables for adjustment in nonexperimental comparative effectiveness research (CER), and uses causal graphs to illustrate the causal network relating treatment to outcome. While selection approaches should be based on an understanding of the causal network representing the common cause pathways between treatment and outcome, the true causal network is rarely known. Therefore, more practical variable selection approaches are described, which are based on background knowledge when the causal structure is only partially known. These approaches include adjustment for all observed pretreatment variables thought to have some connection to the outcome, all known risk factors for the outcome, and all direct causes of the treatment or the outcome. Empirical approaches, such as forward and backward selection and automatic high-dimensional proxy adjustment, are also discussed. As there is a continuum between knowing and not knowing the causal, structural relations of variables, a practical approach to variable selection is recommended, which involves a combination of background knowledge and empirical selection using the high-dimensional approach. The empirical approach could be used to select from a set of a priori variables on the basis of the researcher's knowledge, and to ultimately select those to be included in the analysis. This more limited use of empirically derived variables may reduce confounding while simultaneously reducing the risk of including variables that could increase bias.

Introduction

Nonexperimental studies that compare the effectiveness of treatments are often strongly affected by confounding. Confounding occurs when patients with a higher risk of experiencing the outcome are more likely to receive one treatment over another. For example, consider two drugs used to treat hypertension—calcium channel blockers (CCB) and diuretics. Since many clinicians perceive CCBs as particularly useful in treating high-risk patients with hypertension, patients with a higher risk for experiencing cardiovascular events are more likely to be channeled into the CCB group, thus confounding the relation between antihypertensive treatment and the clinical outcomes of cardiovascular events.[1] The difference in treatment groups is a result of the differing baseline risk for the outcome and the treatment effects (if any). Any attempt to compare the causal effects of CCBs and diuretics on cardiovascular events would require taking patients' underlying risk for cardiovascular events into account through some form of covariate adjustment. The use of statistical methods to make the two treatment groups similar with respect to measured confounders is sometimes called statistical adjustment, control, or conditioning.

The purpose of this chapter is to address the complex issue of selecting variables for adjustment in order to compare the causative effects of treatments. The reader should note that the recommended

variable selection strategies discussed are for nonexperimental causal models and not prediction or classification models, for which approaches may differ. Recommendations for variable selection in this chapter focus primarily on fixed treatment comparisons when employing the so-called "incident user design," which is detailed in chapter 2.

This chapter contains three sections. In the first section, we explain causal graphs and the structural relations of variables. In the second section, we discuss proxy, mismeasured, and unmeasured variables. The third section presents variable selection approaches based on full and partial knowledge of the data generating process as represented in causal graphs. We also discuss approaches to selecting covariates from a high-dimensional set of variables on the basis of statistical association, and suggest how these approaches may be used to complement variable selection based on background knowledge. Ideally, when information is available, causal graph theory would be used to complement any variable selection technique. We provide a separate supplement (supplement 2) on directed acyclic graphs for the more advanced reader.

Causal Models and the Structural Relationship of Variables

This section introduces notation to illustrate basic concepts. Causal graphs are used to represent relationships among variables and to illustrate situations that generate bias and confounding.

Treatment Effects

The goal of comparative effectiveness research (CER) is to determine if a treatment is more effective or safer than another. Treatments should be "well defined," as described in chapter 4, and should represent manipulable units; e.g., drug treatments, guidelines, and devices. Causal graphs are often used to illustrate relationships among variables that lead to confounding and other types of bias. The simple causal graph in Figure 7.1 indicates a randomized trial in which no unmeasured or measured variables influence treatment assignment where A_0 is the assigned treatment at baseline (time zero) and Y_1 is the outcome after followup (time 1). The arrow connecting treatment assignment (A_0) to the outcome (Y_1) indicates that treatment has a causal effect on the outcome. Causal graphs are used to represent the investigator's beliefs about the mechanisms that generated the data. Knowledge of the causal structure that generates the data allows the investigator to better interpret statistical associations observed in the data.

Figure 7.1. Causal graph illustrating a randomized trial where assigned treatment (A_0) has a causal effect on the outcome (Y_1).

Risk Factors

We now let C_0 be one or more baseline covariates measured at time zero. Covariates that are predictive of the outcome but have no influence on treatment status are often referred to as pure risk factors, depicted in Figure 7.2. Conditioning on such risk factors is unnecessary to remove bias but can result in efficiency gains in estimation[2-3] and does not induce bias in regression or propensity score models.[4] Researchers need to be careful not to include variables affected by the outcome, as adjustment for such variables can increase bias.[2] We recommend including risk factors in statistical models to increase the efficiency/precision of an estimated treatment effect without increasing bias.[4]

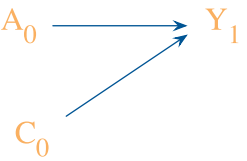

Figure 7.2. Causal graph illustrating a baseline risk factor (C_0) for the outcome (Y_1).

Confounding

The central threat to the validity of nonexperimental CER is confounding. Due to the ways in which providers and patients choose treatments, the treatment groups may not have similar underlying risk for the outcome.

Confounding is often illustrated as a common cause pathway between the treatment and outcome. Measured variables that influence treatment assignment, are predictive of the outcome, and remove confounding when adjusted for are often called confounders. Unmeasured variables on a common cause pathway between treatment and outcome are referred to as unmeasured confounders. For example, in Figure 7.3, unmeasured variables *U1* and *U2* are causes of treatment assignment and outcome. In general, sources of confounding in observational comparative effectiveness studies include provider actions, patient actions, and social and environmental factors. Unmeasured variable *U1* has a measured confounder C_0 that is a proxy for *U1*, such that conditioning on C_0 removes confounding by *U1*, while the unmeasured variable *U2* does not.

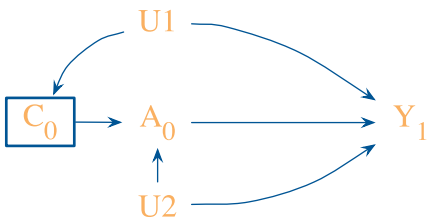

Figure 7.3. A causal graph illustrating confounding from the unmeasured variable *U2*. Conditioning on the measured variable (C_0), as indicated by the box around the variable, removes confounding from *U1*. Measured confounders are often proxies for unmeasurable constructs. For example, family history of heart disease is a measured variable indicating someone's risk for cardiovascular disease (*U1*).

Provider Actions

Confounding by indication: Confounding by indication, also referred to as "channeling bias," is common and often difficult to control in comparative effectiveness studies.[5-9] Prescribers choose treatments for patients who they believe are most likely to benefit or least likely to be harmed. In a now historic example, Huse et al. surveyed United States physicians about their use of various classes of antihypertensive medications and found that physicians were more likely to prescribe CCBs to high-risk patients than for uncomplicated hypertension.[1] Any attempt to compare the safety or effectiveness between CCBs and other classes of antihypertensive medication would need to adequately account for the selective use of CCBs for higher risk patients. If underlying disease severity and prognosis are not precisely measured and correctly modeled, CCBs would appear more harmful or less effective simply because higher risk patients are more likely to receive CCBs. Variables measuring risk for the outcome being investigated need to be adequately measured and modeled to address confounding by indication.

Selective treatment and treatment discontinuation of preventive therapy in frail and very sick patients: Patients who are perceived by a physician to be close to death or who face serious medical problems may be less likely to receive preventative therapies. Similarly, preventative treatment may be discontinued when health deteriorates. This may explain the substantially decreased mortality observed among elderly users of statins and other preventive medications compared with apparently similar nonusers.[10-11] Even though concerns with discontinuation of therapy may be addressed using time-varying measures of treatment, this type of selective discontinuation presents problems when analyzing fixed treatments. For example, when conducting database studies, data are extracted and analyzed on the basis of the specified study period. The more frail elderly who discontinued treatment prior to the study window would appear to have never received treatment.

Patients with certain chronic diseases or patients who take many medications may also have a lower probability of being prescribed a potentially beneficial medication due to concerns regarding drug-drug interactions or metabolic problems.[8] For example, patients with end-stage renal disease are less likely to receive medications for secondary prevention after myocardial infarction.[12] Additionally, in a study assessing the potential for bias in observational studies evaluating use of lipid-lowering agents and mortality risk, the authors found evidence of bias due to an association between noncardiovascular comorbidities and the likelihood of treatment.[11] Due to these findings, researchers have recommended statin use and other chronic therapies as markers for health status in their causal models.[11, 13]

Patient Actions

Healthy user/adherer bias: Patients who initiate a preventive therapy may be more likely than other patients to engage in other healthy, prevention-oriented behaviors. Patients who start a preventive medication may have a disposition that makes them more likely to seek out preventive health care services, exercise regularly, moderate their alcohol consumption, and avoid unsafe and unhealthy activities.[14] Incomplete adjustment for such behaviors representative of specific personality traits can make preventative medications spuriously or more strongly associated with reduced risk of a wide range of adverse health outcomes.

Similar to patients who initiate preventive medications, patients who adhere to treatment may also engage in more healthful behaviors.[14-15] Strong evidence of this "healthy adherer" effect comes from a meta-analysis of randomized controlled trials where good adherence to placebo was found to be associated with mortality benefits and other positive health outcomes.[16] The benefit can be explained by the healthy behaviors of the patients who use the medication as prescribed rather than placebo effects. Treatment adherence is an intermediate variable between treatment assignment and health outcomes. Any attempt to evaluate the effectiveness of treatment rather than the effect of assigned treatment would require time-varying treatment analysis where subjects are censored when treatment is discontinued. Proper adjustment for predictors of treatment discontinuation is required to resolve the selection bias that occurs when conditioning on patients who adhered to assigned treatment.[17-18]

Physician assessment that patients are functionally impaired (defined as having difficulty performing activities of daily living) may also influence their treatment assignment and health outcomes. Functionally impaired patients may be less able to visit a physician or pharmacy; therefore, such patients may be less likely to collect prescriptions and receive preventive health care services.[8] This phenomenon could exaggerate the benefit of prescription medications, vaccines, and screening tests.[8]

Environmental and Social Factors

Access to health care: Within large populations analyzed in multi-use health care databases, patients may vary substantially in their ability to access health care. Patients living in rural areas, for example, may have to drive long distances to receive specialized care.[8] Other patients face different obstacles to accessing health care, such as cultural factors (e.g., trust in the medical system), economic factors (e.g., ability to pay), and institutional factors (e.g., prior authorization programs, restrictive formularies), all of which may have some direct or indirect relation to treatment and study outcomes.[8]

Intermediate Variables

An intermediate variable is generally thought of as a post-treatment variable influenced by treatment that may or may not lie on the causal pathway between the treatment and the outcome. Figures 7.4 and 7.5 illustrate variables affected by treatment. In Figure 7.4, C_0 is a baseline confounder and must be adjusted for, but a subsequent measurement of the variable at a later time (C_1) is on the causal pathway between treatment and outcome. For example, consider the study previously described comparing classes of antihypertensive medications (A_0) on the risk for cardiovascular events (Y_1). The baseline measure of blood pressure is represented by C_0. Blood pressure measured after treatment is initiated, with adequate time for the treatment to reach therapeutic effectiveness and before the outcome assessment, is considered an intermediate variable and is represented by C_1 in Figure 7.4. When the goal of CER is to estimate the total causal effect of the treatment on the outcome, adjustment for variables on the causal pathway between treatment and outcome, such as blood pressure after treatment is initiated (C_1), is unnecessary and is likely to induce bias[2] toward a relative risk of 1.0, though the direction can sometimes be in the opposite direction. The magnitude of bias is greatest if the primary mechanism of action is through the intermediate pathway. Thus, it would be incorrect to adjust for blood pressure measured after the treatment was initiated (C_1), because most of the medication's effects on cardiovascular

disease are mediated through improvements in blood pressure. This kind of overadjustment would mask the antihypertensive effect of the treatment A_0.

Pharmacoepidemiological studies that do not restrict analyses to incident episodes of treatments are subject to this type of overadjustment. Measurement of clinical covariates such as blood pressure at the time of registry enrollment rather than at the time of treatment initiation in an established medication user is such an example. For such patients, a true baseline measurement is unobtainable. The clinical variables for established users at the time of enrollment have already been influenced by investigational treatments and are considered intermediate variables rather than baseline confounders. The ability to adequately adjust for baseline confounders and not intermediate variables is one reason the new user design described in chapter 2 is so highly valued.

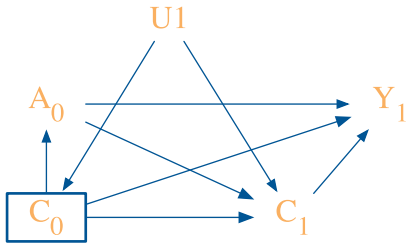

Figure 7.4. A causal graph representing an intermediate causal pathway. Blood pressure after treatment initiation (C_1) is on the causal pathway between antihypertensive treatment (A_0) and cardiovascular events (Y_1). Baseline blood pressure (C_0) is a measured confounder of disease severity ($U1$) and the box around the variable represents adjustment.

Investigators are sometimes interested in separating total causal effects into direct and indirect effects. In mediation analysis, the investigator intentionally measures and adjusts intermediate variables to estimate direct and indirect effects. Mediation analysis requires a stronger set of identifiability assumptions and is discussed in several articles.[19-33]

When conditioning on an intermediate, biases can also arise for "direct effects" if the intermediate is a common effect of the exposure and an unmeasured variable that influences the outcome as in Figure 7.5. The "birth-weight paradox" is one of the better known clinical examples of this phenomenon.[27, 32, 34] Maternal smoking seems to have a protective effect on infant mortality in infants with the lowest birth weight. The seemingly protective effect of maternal smoking is a predictable association produced from conditioning on an intermediate without adequate control for confounding between the low birth weight (intermediate) and infant mortality (outcome). This is illustrated in Figure 7.5. The problem of conditioning on a common effect of two variables will be further discussed below in the section on colliders.

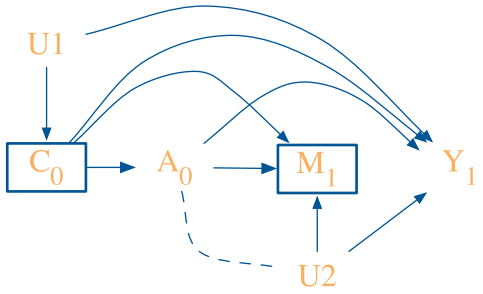

Figure 7.5. A causal diagram illustrating the problem of adjustment for the intermediate variable, low birth weight (M_1), when evaluating the causal effect of maternal smoking (A_0) on infant mortality (Y_1) after adjustment for measured baseline confounders (C_0) between exposure and outcome. Confounding at the intermediate and outcome, birth defects ($U1$), remains unmeasured.

Time-Varying Confounding

The intention-to-treat analogue of a randomized trial, where subjects are assigned to the treatment they are first exposed to regardless of discontinuation or switching treatments, may not be the optimal design for all nonexperimental CER. Researchers interested in comparing adverse effects of medications that are thought to occur only in proximity to using the medication may, for example, want to censor subjects who discontinue treatment. This type of design is described as a "per protocol" analysis. An "as treated" analysis allows subjects to switch treatment groups on the basis of their use of treatment. Both the "as treated" and "per protocol" analysis can be used to evaluate time-varying treatment.

In a nonexperimental setting, time-varying treatments are expected to have time-varying confounders. For example, if we are interested in comparing cardiovascular events between subjects who are completely adherent to CCBs versus completely adherent to diuretics, then we may consider a time-varying treatment design where subjects are censored when they discontinue the treatment to which they were first assigned (as illustrated in Figure 7.6). If joint predictors of compliance and the outcome are present, then some sort of adjustment for the time-varying predictors must be made. Standard adjustment methods may not produce unbiased effects when the predictors of adherence and the outcome are affected by prior adherence, and a newer class of causal effect estimators, such as inverse-probability-of-treatment weights or g-estimation, may be warranted.[18, 35]

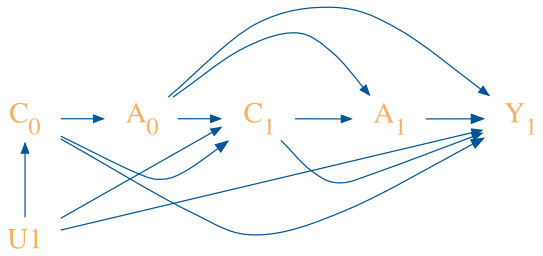

Figure 7.6. A simplified causal graph illustrating adherence to initial antihypertensive therapy as a time-varying treatment (A_0, A_1), joint predictors of treatment adherence and the outcome (C_0, C_1). The unmeasured variable (*U1*) indicates this is a nonexperimental study.

Collider Variables

Colliders are the result of two independent causes having a common effect. When we include a common effect of two independent causes in our statistical model, the previously independent causes become associated, thus opening a backdoor path between the treatment and outcome. This phenomenon can be explained intuitively if we think of two causes (sprinklers being on or it is raining) of a lawn being wet. If we know the lawn is wet, and we know the value of one of the other variables (it is not raining), then we can predict the value of the other variable (the sprinkler must be on). Therefore, conditioning on a common effect induces an association between two previously independent causes, that is, sprinklers being on and rain.

Bias resulting from conditioning on a collider when attempting to remove confounding by covariate adjustment is referred to as *M*-collider bias.[36] Pure pretreatment *M*-type structures that statistically behave like confounders may be rare; nevertheless, any time we condition on a variable that is not a direct cause of either the treatment or outcome but merely associated with the two, we have the potential to introduce *M*-bias.[37]

A hypothetical example of how two independent variables can become conditionally associated and increase bias follows. Consider a highly simplified hypothetical study to compare rates of acute liver failure between new users of CCB and diuretics using administrative data from a distributed network of managed care organizations. As illustrated in Figure 7.7, if some of the managed care organizations had a formulary policy (*U1*) that caused a lower proportion of patients to be initiated on a CCB (A_0), and that same policy reduced the chance of receiving medical treatment for erectile dysfunction (F_0), and patients with a long history of unmeasured alcohol abuse (*U2*) are more likely to receive treatment for erectile dysfunction (F_0), then adjustment for erectile dysfunction treatment may introduce bias by generating an association and opening a backdoor path that did not previously exist between formulary policy (*U1*) and alcohol abuse (*U2*).

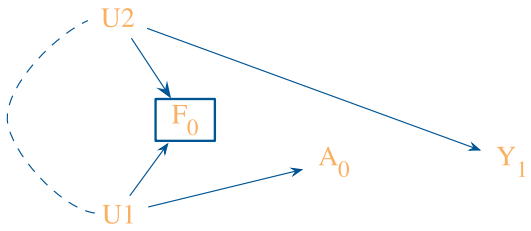

Figure 7.7. Hypothetical causal diagram illustrating *M*-type collider stratification bias. Formulary policy (*U1*) influences treatment with CCB (A_0) and treatment for erectile dysfunction (F_0). Unmeasured alcohol use (*U2*) influences impotence and erectile dysfunction treatment (F_0) and acute liver disease (Y_1). In this example there is no effect of antihypertensive treatment on liver disease, but antihypertensive treatment and liver disease would be associated when adjusting for medical treatment of erectile dysfunction. The box around F_0, represents adjustment and the conditional relationship is represented by the dotted arrow connecting *U1* and *U2*.

Although conditioning on a common effect of two variables can induce an association between two otherwise independent variables, we currently lack many compelling examples of pure *M*-bias for pretreatment covariates. Such structures do, however, arise more commonly in the analysis of social network data.[38] Compelling examples of collider stratification bias (i.e., selection bias) do exist when conditioning on variables affected by treatment (as illustrated in Figure 7.5). Collider stratification bias can give rise to other biases in case-control studies and studies with time-varying treatments and confounding.[39]

Instrumental Variables

An instrumental variable is a pretreatment variable that is a cause of treatment but has no causal association with the outcome other than through its effect on treatment such as Z_0 in Figure 7.8. When treatment has an effect on the outcome, an instrumental variable will be associated with treatment and the outcome, and can thus statistically appear to be a confounder. An instrumental variable will also be associated with the outcome even when conditioning on the treatment variable whenever there is an unmeasured common cause of the treatment on the outcome. It has been established that inclusion in statistical models of variables strongly associated with treatment (A_0) but not independently associated with the outcome (Y_1) will increase the standard error and decrease the precision of the treatment effect.[2, 4, 40-41] It is less well known, however, that the inclusion of such instrumental variables into statistical models intended to remove confounding can increase the bias of an estimated treatment effect. The bias produced by the inclusion of such variables has been termed "*Z*-bias," as *Z* is often used to denote an instrumental variable.[8]

Z-bias arises when the variable set is insufficient to remove all confounding, and for this reason *Z*-bias has been described as bias-amplification.[42-43] Figure 7.8 illustrates a data-generating process where unmeasured confounding exists along with an instrumental variable. In this situation, the variation in treatment (A_0) can be partitioned into three components: the variation explained by the instrument (Z_0), the variation explained by *U1*, and the unexplained variation. The magnitude of unmeasured confounding is determined by the proportion of variation explained by *U1*, along with the association between *U1* and Y_1. When Z_0 is statistically adjusted, one source of variation in A_0 is removed making the variation explained by *U1* a larger proportion of the remaining variation. This is what amplifies the residual confounding bias.[44]

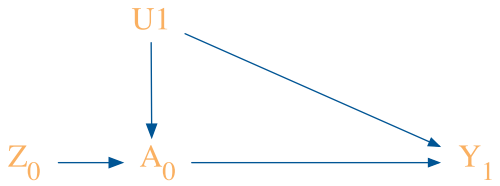

Figure 7.8. Bias is amplified (*Z*-bias) when an instrumental variable (Z_0) is added to a model with unmeasured confounders (*U1*).

Any plausible instrumental variable can potentially introduce *Z*-bias in the presence of uncontrolled confounding. Indication for treatment was found to be a strong instrument[45] and provider and ecologic causes of variation in treatment choice have been proposed as potential instrumental variables that may amplify bias in nonexperimental CER.[8]

A simulation study evaluating the impact of adjusting instruments of varying strength when in the presence of uncontrolled confounding demonstrated that the impact of adjusting instrumental variables was small in certain situations, a result which led the authors to suggest that over-adjustment is less of a concern than under-adjustment. Analytic formulae, on the other hand, indicate that this bias may be quite large, especially when dealing with multiple instruments.[42] We have discussed bias amplification due to adjusting for instrumental variables. The use of instrumental variables, however, can be employed as an alternative strategy to deal with unmeasured confounding.[46] This strategy is discussed in detail in chapter 10.

We have presented multiple types of variable structures, with a focus on variables that either remove or increase bias when adjusted. The dilemma is that many of these variable types statistically behave like confounders, which are the only structural type needing adjustment to estimate the average causal effect of treatment.[47-48] For this reason, researchers should be hesitant to rely on

statistical associations alone to select variables for adjustment. The variable structure must be considered when attempting to remove bias through statistical adjustment.

Proxy, Mismeasured, and Unmeasured Confounders

It is not uncommon for a researcher to be aware of an important confounding variable and to lack data on that variable. A measured proxy can sometimes stand in for an unmeasured confounder. For example, use of oxygen canisters could be a proxy for failing health and functional impairment; use of preventive services, such as flu shot, is sometime thought to serve as a proxy for healthy behavior and treatment adherence. Likewise, important confounders sometimes are measured with error. For example, self-reported body mass index will often be subject to underreporting.

Researchers routinely adjust analyses using proxy confounders and mismeasured confounders. Adjusting for a proxy or mismeasured confounder will reduce bias relative to the unadjusted estimate, provided the effect of the confounder on the treatment and the outcome are "monotonic."[48] In other words, any increase in the confounder should on average always affect treatment in the same direction, and should always affect the outcome in the same direction for both the treated and untreated groups. If an increase in the confounder increased the outcome for the treated group and decreased the outcome for the untreated group, then adjustment for the proxy or mismeasured confounder can potentially increase bias. Unfortunately, there are cases, even when the measurement error of the confounder is nondifferential (i.e., does not depend on treatment or outcome), where adjustment for proxy or mismeasured confounders can increase, rather than decrease, bias.[49]

Another common problem in trying to estimate causal effects is that of unmeasured confounding. Sensitivity analysis techniques have been developed to address misclassified and unmeasured confounding. The reader is referred to chapter 11 for further discussion of sensitivity analyses.

Selection of Variables To Control Confounding

We present two general approaches to selecting variables in order to control confounding in nonexperimental CER. The first approach selects variables on the basis of background knowledge about the relationship of the variable to treatment and outcome. The second approach relies primarily on statistical associations to select variables for control of confounding, using what can be described as high-dimensional automatic variable selection techniques. The use of background knowledge and causal graph theory is strongly recommended when there is sufficient knowledge of the causal structure of the variables. Sufficient knowledge, however, is likely rare when conducting studies across a wide geography and many providers and institutions. For this reason, we also present practical approaches to variable selection that empirically select variables on the basis of statistical associations.

Variable Selection Based on Background Knowledge

Causal Graph Theory

Assuming that a well-defined fixed treatment employing an intention-to-treat paradigm and no set of covariates predicts treatment assignment with 100 percent accuracy, control of confounding is all that is needed to estimate causal effects with nonexperimental data.[47-48] The problem, as described above, is that colliders, intermediate variables, and instruments can all statistically behave like confounders. For this reason, an understanding of the causal structure of variables is required to separate confounders from other potential bias-inducing variables. This dilemma has led many influential epidemiologists to take a strong position for selecting variables for control on the basis of background knowledge of the causal structure connecting treatment to outcome.[50-54]

When sufficient knowledge is available to construct a causal graph, a graphical analysis of the structural basis for evaluating confounding is the most robust approach to selecting variables for adjustment. The goal is to use the graph to identify a sufficient set of variables to achieve unconfoundedness, sometimes also called

conditional exchangeability.[24, 55] The researchers specify background causal assumptions using causal graph criteria (see supplement 2 of this *User's Guide*). If the graph is correct, it can be used to identify a sufficient set of covariates (C) for estimating an effect of treatment (A_0) on the outcome (Y_1). A sufficient set C is observed when no variable in C is a descendant of A_0 and C blocks every open path between A_0 and Y_1 that contains an arrow into A_0. Control of confounding using graphical criteria is usually described as control through the "back-door" criteria, the idea being that variables that influence treatment assignment—that is, variables that have arrows pointing to treatment assignment—provide back-door paths between the A_0 and Y_1. It is the open back-door pathways that generate dependencies between A_0 and Y_1 and can produce spurious associations when no causal effect of A_0 on Y_1 is present, and that alter the magnitude of the association when A_0 causally affects Y_1.

Although it is quite technical, causal graph theory has formalized the theoretical justification for variable selection, added precision to our understanding of bias due to under- and over-adjustment, and unveiled problems with historical notions of statistical confounding. The main limitation of causal graph theory is that it presumes that the causal network is known and that the only unknown is the magnitude of the causal contrast between A_0 and Y_1 being examined. In practice, where observational studies include large multi-use databases spanning vast geographic regions, such complete knowledge of causal networks is unlikely.[56-57]

Since we rarely know the true causal network that represents all common-cause pathways between treatment and outcome, investigators have proposed more practical variable selection approaches based on background knowledge when the causal structure is only partially known. These strategies include adjusting for all observed pretreatment variables thought to have some connection to the outcome,[58] all known risk factors for the outcome,[4, 44, 59] and all direct causes of the treatment or the outcome.[57] The benefits and limitations to each approach to removing confounding are briefly discussed.

Adjustment for All Observed Pretreatment Covariates

Emphasis is often placed on the treatment assignment mechanism and on trying to reconstruct the hypothetical broken randomized experiment that led to the observational data.[58] Propensity score methods are often employed for this purpose and are discussed in chapter 10; they can be used in health care epidemiology to statistically control large numbers of variables when outcomes are infrequent.[60, 61] Propensity scores are the probability of receiving treatment given the set of observed covariates. The probability of treatment is estimated conditional on a set of covariates and the predicted probability is then used as a balancing score or matching variable across treatment groups to estimate the treatment effect.

The greatest importance is often placed on balancing all pretreatment covariates. However, when attempts are made to balance all pretreatment covariates, regardless of their structural form, biases, for example from including strong instruments and colliders, can result,[37, 57, 62] though, as noted above, in practice, pretreatment colliders are likely rarer than ordinary confounding variables.

Adjustment for All Possible Risk Factors for the Outcome

Confounding pathways require common cause structures between the outcome and treatment. A common strategy for removing confounding without incidentally including strong instruments and colliders is to include in propensity score models only variables thought to be direct causes of the outcome, that is, risk factors.[4, 59, 63] This approach requires only background knowledge of causes of the outcome, and it does not require an understanding of the treatment assignment mechanism or how variables that influence treatment are related to risk factors for the outcome. This strategy, however, may fail to include measured variables that predict treatment assignment but have an unmeasured ancestor that is an outcome risk factor ($A_0 \leftarrow C_0 \leftarrow U1 \rightarrow Y_1$) as illustrated in Figure 3.[57]

Disjunctive Cause Criterion

The main practical use of causal graphs is to ensure adjustment for confounders and avoid adjusting for known colliders.[51] In practice, one only needs to partly know the causal structure of variables relating treatment to the outcome. The disjunctive cause criterion is a formal statement of the conditions in which variable selection based on partial knowledge of the causal structure can remove confounding.[57] It states that all observed variables that are a cause of treatment, a cause of outcome, or a cause of both should be included for statistical adjustment. It can be shown that when any subset of observed variables is sufficient to control confounding, the set obtained by applying the disjunctive cause criteria will also constitute a sufficient set.[57] This approach requires more knowledge of the variables' relationship to the treatment and outcome using all pretreatment covariates, or all risk factors, but less knowledge than the back-door path criterion.

Whenever there exists some set of observed variables that block all back-door paths (even if the researcher does not know which subset this is), the disjunctive cause criterion when applied correctly by the investigators will identify a set of variables that also blocks all back-door paths. The other variable selection criteria based on all pretreatment covariates and risk factors do not have this property.[57] The approach performs well when the measured variables include some sufficient set, but presents problems when unmeasured confounding remains. In this case, conditioning on an instrument can amplify the bias due to unmeasured confounding. Thus, in practice, known instruments should be excluded before applying the criterion. The best approach to variable selection is less clear when unmeasured confounding may remain after statistical adjustment for measured variables, which is often expected in nonexperimental CER. In this case, every variable selection approach will result in bias. The focus would then be on minimizing bias, which requires thoughtful consideration of the tradeoff between over- and underadjustment. Strong arguments exist for error on the side of overadjustment (adjusting for instruments and colliders) rather than failing to adjust for measured confounders (underadjustment).[36, 44] Nevertheless, adjustments for instrumental variables have been found to amplify bias in practice.[45]

Empirical Variable Selection Approaches

Historically, data for nonexperimental studies was primarily collected prospectively, and thoughtful planning was needed to ensure complete measurement of all important study variables. We now live in an era where every interaction between the patient and the health care system produces hundreds, if not thousands, of data points that are recorded for clinical and administrative purposes.[64] These large multi-use data sources are highly dimensional in that every disease, medication, laboratory result, and procedure code, along with any electronically accessible narrative statements, can be treated as variables.

The new challenge to the researcher is to select a set of variables from this high-dimensional space that characterizes the patient's baseline status at the time of treatment selection to enable identification of causal effects, or that at least produces the least biased estimates. Advances in computer performance and the availability of high-dimensional data have provided unprecedented opportunities to use data empirically to "learn" associational relationships. Empiric variable selection techniques include identifying a subset of variables of statistical associations with the treatment and/or outcome from the original set on the basis of background knowledge of the relationship with treatment and/or outcome, as well as methods that are considered fully automated, where all variables are initially selected on the basis of statistical associations.

Forward and Backward Selection Procedures

When using traditional regression it is not uncommon to use, for the purposes of covariate selection, what are sometimes called forward and backward selection procedures. Forward selection procedures begin with an empty set of covariates and then consider whether for each covariate, the covariate is associated with the outcome conditional on treatment (usually using a p-value cutoff in a regression model of 0.05 or 0.10). The variable that is most strongly associated with outcome (based on having the smallest p-value

below the cutoff) is then added to the collection of variables for which control will be made. Then the process begins again, and one considers whether each covariate is associated with the outcome conditional on the treatment and the covariate already selected; the next covariate that is most strongly associated is again added to the list. The process repeats until all remaining covariates are independent of the outcome conditional on the treatment and the covariates that have been previously selected for control.

Backward selection begins with all covariates in the model; then the investigator considers whether, for each covariate, that covariate is independent of the outcome conditional on the treatment and all other covariates (generally using a p-value cutoff in a regression model of 0.05 or 0.10). The covariate with the largest p-value above the cutoff is then discarded from the list of covariates for which control is made. The process begins again, and the investigator considers whether, for each covariate, that covariate is independent of the outcome, conditional on the treatment and the other covariates not yet discarded; the next covariate with the weakest association with the outcome based on p-value is again discarded. The process repeats itself until all variables still in the list are associated with the outcome conditional on the treatment and the other covariates that have not been discarded.

Provided that the original set of covariates with which one begins suffices for unconfoundedness of treatment effects estimates, then if the backward selection process correctly discards variables that are independent of the outcome conditional on the treatment and other covariates, the final set of covariates selected by the backwards selection procedure will also yield a set of covariates that suffices for conditional exchangeability.[57] Likewise, under an additional assumption of "faithfulness,"[57] the forward selection procedure will identify a set of covariates that suffices for unconfoundedness provided that the original set of covariates with which one begins suffices to achieve unconfoundedness and that the forward selection process correctly identifies the variables that are and are not independent of the outcome conditional on the treatment and other covariates. The forward and backward procedures can thus be useful for covariate reduction, but both of them suffer from the need to specify a set of covariates to begin with that suffice for unconfoundedness. Thus, even if an investigator intends to employ forward or backward selection procedures for covariate reduction, other approaches will be needed to decide on what set of covariates these forward and backward procedures should begin with. Moreover, when the initial set of covariates does not suffice for unconfoundedness, it is not clear how forward and backward selection procedures will perform. Variable selection procedures also suffer from the fact that estimates about treatment effects are made after having already used the data to decide on covariates.

Similar but more sophisticated approaches using machine learning algorithms such as boosting, random forest, and other ensemble methods have become increasingly common, as have sparsity-based methods such as LASSO, in dealing with high-dimensional data.[65] All of these empirically driven methods are limited, however, in that they are in general unable to distinguish between instruments, colliders, and intermediates on the one hand and genuine confounders on the other. Such differentiation needs to be made a priori on substantive grounds.

Automatic High-Dimensional "Proxy" Adjustment

In an attempt to capture important proxies for unmeasured confounders, Schneeweiss and colleagues proposed an algorithm that creates a very large set of empirically defined variables from health care utilization data.[56] The created variables capture the frequency of codes for procedures, diagnoses, and medication fills during a pre-exposure period. The variables created by the algorithm are required to have a minimum prevalence in the source population and to have some marginal association with both treatment and outcome. After they are defined, the variables can be entered into a propensity score model. In several example studies where the true effect of a treatment was approximately known from randomized controlled trials, the algorithm appeared to perform as well as or better than approaches based on simply adjusting for an a priori set of variables.[45, 66] By defining variables prior to treatment, propensity score methods will

not "over-adjust" by including causal intermediates. Using statistical associations to select potential confounders can result in selection and adjustment of colliders and instruments. Therefore, the analyst should attempt to remove such variables from the set of identified variables. For example, variables that are strong predictors of treatment but have no obvious relation to the outcome should be considered potential sources of *Z*-bias.

A Practical Approach Combining Causal Analysis With Empirical Selection

There is a continuum between knowing and not knowing the causal, structural relations of variables. We suggest that a practical approach to variable selection may involve a combination of (1) a priori variable selection based on the researcher's knowledge of causal relationships together with (2) empirical selection using the high-dimensional approach described above.[8] The empirical approach could be used to select from a set of a priori variables on the basis of the researcher's knowledge, and to ultimately select those to be included in the analysis. This more limited use of empirically derived variables may reduce confounding while simultaneously reducing the risk of including variables that could increase bias.

Conclusion

In practice, the particular approach that one adopts for observational research will depend on the researcher's knowledge, the data quality, and the number of covariates. A deep understanding of the specific clinical and public health risks and opportunities that lie behind the research question often drives these decisions.

Regardless of the strategy employed, researchers should clearly describe how variables are measured and provide a rationale for a priori selection of potential confounders, ideally in the form of a causal graph. If the researchers decide to further eliminate variables using an empiric variable selection technique, then they should present both models and describe what criteria were used to determine inclusion and exclusion. Researchers should consider whether or not they believe adequate measurement is available in the dataset when employing a specific variable selection strategy. In addition, all variables included for adjustment should be listed in the manuscript or final report. When empirical selection procedures are newly developed or modified, researchers are encouraged to make the protocol and code publicly available to improve transparency and reproducibility.

Even when researchers use the methods we describe in this chapter, confounding can persist. Sensitivity analysis techniques are useful for assessing residual confounding resulting from unmeasured and imperfectly measured variables.[67-75] Sensitivity analysis techniques assess the extent to which an unmeasured variable would have to be related to the treatment and outcome of interest in order to substantially change the conclusions drawn about causal effects. We refer the reader to chapter 11 for discussion of sensitivity analysis techniques.

Checklist: Guidance and key considerations for covariate selection in CER protocols		
Guidance	**Key Considerations**	**Check**
Describe the data source(s) that will be used to identify important covariates.	– Provide information about the source(s) of data for key covariates, acknowledging the strengths and weaknesses of the data source (e.g., administrative claims, EMRs, chart review, patient self-report) for measuring each type of covariate.	☐
Discuss the potential for unmeasured confounding and misclassification.	– Discuss the potential impact of unmeasured confounders and misclassification or measurement error. – Propose specific formal sensitivity analysis of the impact of unmeasured confounders or misclassified variables.	☐
Describe the approach to be used to select covariates for statistical models.	– Discuss approaches based on background knowledge (e.g., selection of all hypothesized common causes, disjunctive cause criterion, directed acyclic graphs, or selection of all variables thought to be risk factors for the outcome. – Describe model reduction techniques to be used (e.g., forward or backward selection). – Describe empirical variable selection techniques and how variables were removed from consideration when they were thought to be bias-inducing rather than bias-reducing variables.	☐

References

1. Huse DM, Roht LH, Hartz SC. Selective use of calcium channel blockers to treat high-risk hypertensive patients. Pharmacoepidemiol Drug Saf. 2000;9(1):1-9.
2. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiol. 2009;20(4):488-95.
3. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol. 1986;123(3): 392-402.
4. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413-9.
6. Greenland S, Neutra R. Control of confounding in the assessment of medical technology. Int J Epidemiol. 1980;9(4):361-7.
7. Blais L, Ernst P, Suissa S. Confounding by indication and channeling over time: the risks of beta 2-agonists. Am J Epidemiol. 1996;144(12):1161-9.
8. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010;48(6 Suppl):S114-20.
9. Joffe MM. Confounding by indication: the case of calcium channel blockers. Pharmacoepidemiol Drug Saf. 2000;9(1):37-41.
10. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006;98(3):253-9.
11. Glynn RJ, Schneeweiss S, Wang PS, et al. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. J Clin Epidemiol. 2006;59(8):819-28.
12. Winkelmayer WC, Levin R, Setoguchi S. Associations of kidney function with cardiovascular medication use after myocardial infarction. Clin J Am Soc Nephrol. 2008;3(5):1415-22.
13. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. Epidemiol. 2001;12(6):682-9.

14. Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. Am J Epidemiol. 2007;166(3):348-54.
15. White HD. Adherence and outcomes: it's more than taking the pills. Lancet. 2005;366(9502):1989-91.
16. Simpson SH, Eurich DR, Majumdar SR, et al. A meta-analysis of the association between adherence to drug therapy and mortality. BMJ. 2006;333(7557):15.
17. Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. ClinTrials. 2012; 9(1):48-55.
18. Toh S, Hernan MA. Causal inference from longitudinal studies with baseline randomization. Int J Biostat. 2008;4(1):Article22.
19. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010;172(12):1339-48.
20. VanderWeele TJ. Mediation and mechanism. Eur J Epidemiol. 2009;24(5):217-24.
21. Vanderweele TJ. Causal mediation analysis with survival data. Epidemiol. 2011;22(4):582-5.
22. Vanderweele TJ. Subtleties of explanatory language: what is meant by "mediation"? Eur J Epidemiol. 2011;26(5):343-6.
23. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiol. 2010;21(4):540-51.
24. Pearl J. An introduction to causal inference. Int J Biostat. 2010;6(2):Article7.
25. Moodie EE, Stephens DA. Using directed acyclic graphs to detect limitations of traditional regression in longitudinal studies. Int J Public Health. 2010;55(6):701-3.
26. Hafeman DM. Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. Am J Epidemiol. 2011;174(6):710-7.
27. Whitcomb BW, Schisterman EF, Perkins NJ, et al. Quantification of collider-stratification bias and the birthweight paradox. Paediatr Perinat Epidemiol. 2009;23(5):394-402.
28. Shpitser I, Vanderweele TJ. A complete graphical criterion for the adjustment formula in mediation analysis. Int J Biostat. 2011;7(1):16.
29. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiol. 1992;3(2):143-55.
30. Pearl J. Causal inference from indirect experiments. Artificial intelligence in medicine 1995;7(6):561-82.
31. Young GP, St John DJ, Cole SR, et al. Prescreening evaluation of a brush-based faecal immunochemical test for haemoglobin. J Med Screen. 2003;10(3):123-8.
32. Vanderweele TJ, Mumford SL, Schisterman EF. Conditioning on intermediates in perinatal epidemiology. Epidemiol. 2012;23(1):1-9.
33. Robins J. The control of confounding by intermediate variables. Stat Med. 1989;8(6): 679-701.
34. Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight "paradox" uncovered? Am J Epidemiol. 2006;164(11):1115-20.
35. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2009;28(12):1725-38.
36. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiol. 2003;14(3):300-6.
37. Pearl J. Myth, confusion, and science of causal analysis [Unpublished Manuscript]. Los Angeles, CA: University of California; 2009. http://ftp.cs.ucla.edu/pub/stat_ser/r348-warning.pdf. Accessed March 29, 2012.
38. Shalizi C, Thomas A. Homophily and contagion are generically confounded in observational social network studies. Sociol Methods Res. 2011;40(2):211-39.
39. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiol. 2004;15(5):615-25.
40. Robinson LD, Jewell NP. Covariate adjustment. Biometrics. 1991;47(1):342-3.
41. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078-94.
42. Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol. 2011;174(11):1223-7.

43. Wooldridge J. Should instrumental variables be used as matching variables? [Unpublished Manuscript]. East Lansing, MI: Michigan State University; 2009. https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf. Accessed March 29, 2012.

44. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174(11):1213-22.

45. Patrick AR, Schneeweiss S, Brookhart MA, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiol Drug Saf. 2011;20(6):551-9.

46. Angrist JG, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91:28.

47. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc. 2005;100(469):10.

48. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiol. 2008;19(6):766-79.

49. Brenner H. Bias due to non-differential misclassification of polytomous confounders. J Clin Epidemiol. 1993;46(1):57-63.

50. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiol. 1999;10(1):37-48.

51. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol. 2002;155(2):176-84.

52. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiol. 2001;12(3):313-20.

53. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82:669-88.

54. Glymour MM, Weuve J, Chen JT. Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias. Neuropsychology Rev. 2008;18(3):194-213.

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

56. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiol. 2009;20(4):512-22.

57. Vanderweele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011;67(4):1406-13.

58. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26(1):20-36.

59. Hill J. Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008;27(12): 2055-61; discussion 2066-9.

60. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Int Med.1997;127(8 Pt 2):757-63.

61. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol.2003;158(3):280-7.

62. Pearl J. Remarks on the method of propensity score. Stat Med. 2009;28:1415-24.

63. Myers JA, Rassen JA, Gagne JJ, et al. Myers et al. respond to "understanding bias amplification." Am J Epidemiol. 2011;174(11):1228-9.

64. D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. Am J Med. 2010;123(12 Suppl 1):e32-7.

65. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267-288.

66. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol. 2011;173(12):1404-13.

67. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiol. 2011;22(1):42-52.

68. Vanderweele TJ. Sensitivity analysis: distributional assumptions and confounding assumptions. Biometrics. 2008;64(2):645-9.

69. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf. 2006;15(5): 291-303.

70. Rosenbaum PR. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics. 2007;63(2):456-64.

71. Rosenbaum PR. Sensitivity analysis for matched case-control studies. Biometrics. 1991;47(1):87-100.

72. Greenland S. Useful methods for sensitivity analysis of observational studies. Biometrics. 1999;55(3):990-1.

73. Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996. 25(6):1107-16.

74. Brumback BA, Hernán MA, Haneuse SJ, et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med. 2004;23(5):749-67.

75. Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Ann Epidemiol. 2008;18(8):637-46.