ANNUAL REVIEWS

Annual Review of Clinical Psychology

Accounting for Confounding in Observational Studies

Brian M. D'Onofrio,[1,2] Arvid Sjölander,[2] Benjamin B. Lahey,[3] Paul Lichtenstein,[2] and A. Sara Öberg[2,4]

[1]Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana 47405, USA; email: bmdonofr@indiana.edu

[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden; email: arvid.sjolander@ki.se, paul.lichtenstein@ki.se, oberg@hsph.harvard.edu

[3]Departments of Health Studies and Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois 60637, USA; email: blahey@health.bsd.uchicago.edu

[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA

Annu. Rev. Clin. Psychol. 2020. 16:25–48

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

https://doi.org/10.1146/annurev-clinpsy-032816-045030

Copyright © 2020 by Annual Reviews.
All rights reserved

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

causation, confounding, causal diagram, propensity scores, natural experiments, quasi-experiments

Abstract

The goal of this review is to enable clinical psychology researchers to more rigorously test competing hypotheses when studying risk factors in observational studies. We argue that there is a critical need for researchers to leverage recent advances in epidemiology/biostatistics related to causal inference and to use innovative approaches to address a key limitation of observational research: the need to account for confounding. We first review theoretical issues related to the study of causation, how causal diagrams can facilitate the identification and testing of competing hypotheses, and the current limitations of observational research in the field. We then describe two broad approaches that help account for confounding: analytic approaches that account for measured traits and designs that account for unmeasured factors. We provide descriptions of several such approaches and highlight their strengths and limitations, particularly as they relate to the etiology and treatment of behavioral health problems.

Contents

1. OVERVIEW

A fundamental aim of clinical psychology is to improve our understanding of the etiology, treatment, and prevention of behavioral health problems. To do this, the field must integrate advances from other disciplines. In this review, we argue that there is a critical need for the field of clinical psychology to leverage recent advances in epidemiology/biostatistics related to causal inference and to use innovative designs to address a key limitation of observational research: the need to account for confounding. This certainly applies to clinical scientists who design and/or analyze large cohort studies, but understanding and incorporating a more formal approach to causal inference is also essential for other clinical psychology research, including in the fields of experimental psychopathology and clinical neuroscience. The major goal of this review is to enable researchers to identify and more rigorously test competing hypotheses when studying risk factors in observational studies that do not involve random assignment by bridging the fields of clinical psychology and epidemiology/biostatistics.

Confounding: the spurious or noncausal association between two variables that arises because both variables are caused by another factor(s)

Observational studies: studies that do not use random assignment to examine causal effects

Bridging these fields can be difficult for several reasons. First, the fields frequently use different terms for the same issue, and sometimes the same terms are used to represent very different concepts. Second, the types of variables that the fields have typically explored are different. Epidemiology has historically focused more on categorical (e.g., binary) variables, whereas psychology has primarily focused on continuous measures. This may seem like a minor detail, but the difference greatly influences the type of methodological training students receive. Third, historical differences in the general levels of analysis can further complicate communication. Yet, we believe that bridging the gap can greatly help the field of clinical psychology—a goal that is consistent with other calls in psychology (Rohrer 2018) and epidemiology (Greenland et al. 1999) as well as related fields (Elwert & Winship 2014).

To achieve our goal, this review covers two main topics. First, we introduce advances in the study of causal inference with a focus on the critical necessity of rigorously exploring alternative explanations—in particular, the need to identify and account for confounding. Our hope is to provide psychologists with an introduction to key theoretical issues in the study of causation, the use of graphical tools, and an appraisal of key limitations of observational research in clinical psychology. Second, we review several (*a*) analytical approaches and (*b*) research designs that help address confounding and illustrate their use in clinical psychology. While we cannot cover all possible designs, we provide several exemplars to illustrate the importance of using advanced approaches to test competing hypotheses, and we note the limitations inherent in each approach. We conclude the review by providing several summary points and future directions.

2. CAUSAL INFERENCE

Testing causal hypotheses using observational data in clinical psychology requires an understanding of several issues from philosophy of science (e.g., O'Donahue 2013, Shadish et al. 2002). To start, it requires a formal definition of causality and an informed view of how scientists investigate, explain, and acquire knowledge.

2.1. Counterfactual Reasoning

Clinical psychology is interested in identifying causal risk factors for behaviors that are complex and multifactorial. As such, researchers will not find "big, simple explanations" (Kendler 2005, p. 434; see also Kendler 2019). Unfortunately, causal reasoning, particularly when studying risk factors with such putative small effects, is often poorly motivated and sometimes relies on "the dubious value of causal criteria" (Rothman & Greenland 2005, p. S147). In recent decades, there has been a rapid development toward a formal theory of causal inference (Pearl 2009).

There is a growing consensus for using the counterfactual model to understand causal effects (for an introduction geared to those in the social sciences, see Shadish et al. 2002). For a given individual, a causal effect is defined as a contrast between the outcome had the person been exposed to a risk factor (e.g., received a treatment) at a particular point in time and the outcome had the individual not been exposed (e.g., did not receive a treatment) at that point in time. In practice, though, we can only observe one of these scenarios: the one that actually happened. The other scenario remains unobserved, or counterfactual, and therefore an individual causal effect typically cannot be identified.

At the population level, a causal effect translates to the average outcome had everyone been exposed relative to the average outcome had everyone not been exposed. Yet, we may only observe the counterfactual outcome under exposure among those actually exposed, and vice versa. The conceptual difference between (*a*) comparing separate subsets of individuals defined by their actual exposure level and (*b*) comparing the same set of individuals under different counterfactual exposure scenarios explains the well-known dictum that association is not causation. As a result, "two central tasks in experimental design are creating a high-quality but necessarily imperfect source of counterfactual inference and understanding how this source differs from the treatment condition" (Shadish et al. 2002, pp. 5–6).

We want to emphasize two implications of the counterfactual approach to understanding causal effects for clinical psychologists. First, for the counterfactual contrast to be well defined, researchers must strive for a clear understanding of the causal question (e.g., what hypothetical intervention does it correspond to?). Second, when comparing the outcome among those who were actually exposed and those who were not exposed, it is key that researchers make sure the

groups do not differ in their risk of the outcome for reasons other than the risk factor of interest, or the (factual) association between the risk factor and the outcome will not represent the (counterfactual) causal effect.

Bias: systematic error in the estimation of a causal effect

2.2. Bias and Internal Validity

Underlying differences in the risk of the outcome between those who were exposed and unexposed lead to bias, which is systematic error in the estimation of a causal effect. In this review, we show how bias arises from noncausal pathways between a risk factor and an outcome of interest to create spurious (i.e., noncausal) associations. We particularly focus on how researchers can identify and account for the most common source of bias in observational studies: confounding. This review, thus, is concerned with improving internal validity: the degree to which the association between a risk factor and an outcome reflects a causal effect. We stress that in observational studies, causal effects and confounding are competing, though not mutually exclusive, hypotheses about the processes that give rise to an association between a risk factor and an outcome. To what degree does the observed association reflect a potential causal effect of the risk factor on the outcome versus a noncausal association created by confounding factors?

2.3. Epistemology

Some basic tenets from scientific epistemology are relevant. First, researchers need to formulate their hypotheses so that they can be vigorously tested (e.g., Popper 1962). Second, advancing our understanding of causality from observational studies requires that researchers identify and then explicitly rule out competing hypotheses—specifically, noncausal explanations of the observed associations—rather than conduct studies that are solely designed or implemented to provide confirming evidence of a particular hypothesis (e.g., Platt 1964). Third, researchers need to search out disconfirming evidence for a particular theory and conduct severe tests of a theory (e.g., Mayo 1996). Establishing the veracity of a claim of a causal effect with observational data, therefore, requires researchers to specifically account for alternative, noncausal explanations for why a risk factor is associated with an outcome.

Whereas many readers may find this brief exposition in epistemology to be quite basic and consequently unremarkable, we believe that one of the major limitations of observational research in clinical psychology is a fundamental failure to identify competing explanations for observed associations between risk factors and outcomes. Psychological research is too often muddied by confirmation bias (e.g., a priori assumptions about causal effects) that can negatively influence all aspects of observational research from study design and analysis to interpretation and dissemination of findings. For example, psychologists have historically made strong causal claims about psychosocial risk factors in observational data because of a priori assumptions of environmental effects (Rutter 2000).

In this article, we review how advanced statistical approaches and designs can help account for confounding in observational studies, but we stress that the ultimate success rests first and foremost on researchers identifying all plausible competing explanations for why a risk factor would be associated with an outcome. Only then can researchers design a series of tests to account for competing hypotheses in an iterative fashion (Platt 1964). We believe that the use of causal diagrams can be instrumental in guiding researchers in this endeavor.

2.4. Causal Diagrams

To understand the challenges to causal inference, it is vital that researchers make their hypotheses, knowledge, and assumptions explicit. Causal diagrams help clarify causal questions and identify

whether and how they may be tested. They also allow researchers to describe, explain, and classify sources of bias, including confounding. For an introduction to causal diagrams for beginners, readers are referred to Rohrer (2018) and Elwert & Winship (2014), while initiated readers may turn to the work of Pearl (2009) for a formal exposition.

Collider: a variable that is caused by two other variables

Marginal association: the unadjusted relation between two variables (i.e., the raw/crude association)

Variables in a causal diagram are linked by directed arrows, which represent a possible causal effect from one variable to the other. Because a cause must precede its effect, the graph is acyclic; following the direction of the arrows, one can never end up where one started. Some readers may be accustomed to the path diagrams typically used in structural equation modeling. While these bear many similarities to causal diagrams, what characterizes the latter is that they encode all assumptions about the structural (i.e., causal) relationship between variables and include rules to help identify the testable implications of those assumptions (Pearl 2009).

Here, we review the rules for causal diagrams that are particularly relevant for understanding how to distinguish causal effects from confounding. Paths are connections between variables denoted by a single arrow or across multiple arrows. Two variables are associated if there is at least one open path (indicated in the diagrams by an arrow or series of arrows) between them. When researchers make no adjustments, all paths are open, except those paths on which two arrows meet on a single variable, making it a common effect, which is also referred to as a collider. A path that follows the direction of arrows represents a causal pathway, whereas a path that at some point goes against the direction of arrows represents a noncausal pathway.

Figure 1 illustrates these key points. In **Figure 1*a***, the question is whether the risk factor X has a causal effect on outcome Y. The causal diagram includes Z, which represents the set of all factors that cause both X and Y, which are common causes or confounding factors. According to **Figure 1*a***, X and Y are associated through two possible paths: one through the arrow from X to Y (a hypothesized causal path) and the other from X to Z to Y (a hypothesized noncausal path). Confounding, thus, is the spurious or noncausal association that arises because the risk factor (X) and outcome (Y) share common causes (Z). Based on **Figure 1*a***, the crude or unadjusted estimate of the association between X and Y (sometimes referred to in epidemiology as a marginal association) is biased because of confounding. Confounding is the most common source of bias in observational studies because among all the factors that determine who becomes exposed versus not exposed, some are likely to also be risk factors for the outcome.

Causal diagrams help researchers test causal hypotheses because applying the criteria of graph theory allows a priori expectations of whether a risk factor will appear to be associated with an outcome after one or more of the other confounding variables in the diagram are held constant. Estimating an association while holding other variables constant produces an adjusted, or

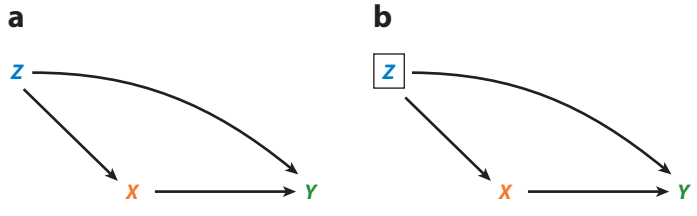

Figure 1

Causal diagram illustrating confounding. X represents a risk factor, and Y represents an outcome. Z represents all common causes of both X and Y. (*a*) The unadjusted association between X and Y would not reflect a causal effect because the variables are also associated through Z. (*b*) Conditioning on Z, noted by a box, blocks the path from X and Y via Z. A conditional association between X and Y (holding constant Z) would reflect the true causal effect.

Conditional association: the adjusted relation between two variables obtained by holding constant another variable(s)

conditional, association. It follows that conditioning on (i.e., adjusting or controlling for) a variable means that it is held constant (so it cannot have any influence) through various methods (see Section 3 for more on statistical methods and designs that enable researchers to condition on a variable). In these diagrams, conditioning on a variable is represented graphically by enclosing the variable in a box. Adjusting for a variable will block all causal and noncausal paths that go through the variable, except if the variable is a collider. For example, conditioning on all of the common causes (e.g., by adjusting for precise measures of all of the factors that *Z* represents), as illustrated by the box around *Z* in **Figure 1*b***, will block the path from *X* to *Y* via *Z*. In Section 2.5.4, we stress the challenges of identifying and accounting for all of these factors by distinguishing between measured and unmeasured confounding. Because the only remaining open path between *X* and *Y* is the direct arrow from *X* to *Y*, the adjusted (or conditional) association between *X* and *Y* in this scenario would reflect the causal effect of *X* on *Y*. Making causal claims thus requires blocking all noncausal paths between the risk factor and outcome of interest.

To further illustrate the use of a causal diagram, consider a true experiment involving randomization. In randomized controlled trials (RCTs), the purpose of randomization is to make sure that there are no systematic differences in preintervention status between the two experimental groups. In the causal diagram in **Figure 1**, randomization theoretically breaks the influence of common causes by removing the arrow from *Z* to *X*. All of the measured and unmeasured confounding factors represented by *Z* are still associated with *Y*, but their relation to *X* has been blocked by design (i.e., through randomization). As such, the observed association between the treatment *X* and outcome *Y* will reflect the causal effect. This ability to avoid confounding due to *Z* is undoubtedly a tremendous strength of RCTs, which are commonly referred to as the gold standard for causal inference. Still, RCTs have other challenges and limitations (for reviews, see Deaton & Cartwright 2018, Rawlins 2008, West 2009). For example, RCTs are frequently too small to study rare-but-serious outcomes, rely on a number of assumptions, and have limited ability to generalize to important patients, treatments, and settings. Finally, several important risk factors do not lend themselves to randomization for ethical or feasibility reasons. Consequently, there are critical questions in clinical psychology that cannot be answered by RCTs and thus require observational studies.

Ultimately, a causal diagram should graphically represent all confounding pathways between the putative causal risk factor and the outcome under study. If a confounding path is left out, then the diagram may give the false impression that the observed association, adjusted for all confounding factors identified in the diagram, can be interpreted as a causal effect. This is why the validity of a causal diagram and the inference made from it rest heavily on subject-matter understanding (Robins 2001).

2.5. Sources of Confounding

The common causes of *X* and *Y* represented by *Z* in **Figure 1** can be of any type. Here, we discuss two possible sources of confounding that have often been neglected in clinical psychology research: previous behavior and genetic factors (Rutter et al. 2001). We do so to provide additional examples of how causal diagrams represent competing (i.e., causal and noncausal) hypotheses for an observed association between a risk factor and an outcome.

2.5.1. Confounding from previous behavior. When studying putative environmental factors, "there must be a strategy to differentiate environmental effects on the person from the effects of individuals on their environments" (Rutter et al. 2001, p. 297). Although dynamic scenarios (where the risk factor and outcome can vary over time and influence each other) are beyond the scope of

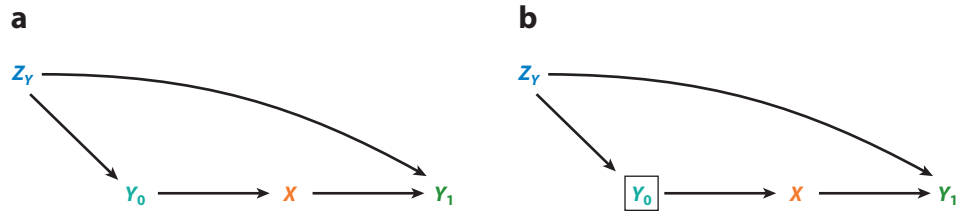

Figure 2

Causal diagram illustrating confounding by previous behavior. X represents a risk factor, and Y_1 represents a behavioral outcome after the exposure. Y_0 represents the behavior of interest measured before the exposure. Z_Y represents all of the (unknown) factors that influence Y over time. (*a*) The unadjusted association between X and Y_1 would not estimate the causal effect because the variables are also associated through the path over Y_0 and Z_Y. (*b*) Conditioning on Y_0, noted by a box, blocks the path from X and Y_1 via Y_0 and Z_Y. A conditional association between X and Y (holding constant Y_0) would reflect the true causal effect.

this review, we present here a simple example in which only the outcome varies with time. The diagram in **Figure 2** illustrates how a risk factor X is hypothesized to have a causal influence on a behavior Y_1, but the potential presence of the same or similar behavior Y_0 before the exposure would also influence the likelihood of the risk factor (X) occurring. The likelihood that an individual exhibits the behavior Y is also under the influence of various underlying factors captured in Z_Y. In **Figure 2*a***, the unadjusted association between X and Y_1 would not represent a causal effect because there is an open (i.e., uncontrolled) noncausal pathway linking the two variables (i.e., from X to Y_0 to Z_Y to Y_1). For this specific scenario, a study design in which researchers could control for pre-exposure behavior Y_0 (illustrated in **Figure 2*b***) would allow the association between X and Y_1 to represent a causal effect because the noncausal path would be blocked. **Figure 2** thus illustrates why clinical psychology research must consider such confounding. The classic text by Bell (1968), for example, highlights that research on socialization from parents to their children must consider the children's effects on the parents. Cross-sectional designs in which X and Y are measured at the same time cannot distinguish such potential reverse effects. Given the complexity of dynamic settings, each scenario requires careful identification of the causal structure and an appropriate approach to handle it (e.g., Howards et al. 2007).

2.5.2. Genetic confounding. Behavior genetic research has shown that the influence of genetic factors on individual differences in behavior is pervasive (Polderman et al. 2015, Turkheimer 2000). Furthermore, genetic factors regularly do not solely influence individual traits but, rather, have pleiotropic effects, so genetic factors are associated with numerous outcomes (Bulik-Sullivan et al. 2015). Genetic effects extend to exposure to measures that are considered to be environmental; this correlation is referred to as gene–environment correlation (Jaffee & Price 2012, Plomin & Bergeman 1991). Genetic variants can become correlated with environmental risk factors in three ways. Passive gene–environment correlation occurs when parents' genetic factors influence their childrearing and are also passed down to their offspring (i.e., characteristics of the children do not affect their own exposure). Active gene–environment correlation, rather, occurs when an individual's genetically influenced characteristics influence their "choice" of (i.e., selection into) environments. Evocative gene–environment correlation occurs when genetically influenced behaviors evoke changes in the individual's environmental risk factors.

As a consequence of gene–environment correlation, genetic factors could be a common cause of a putative causal risk factor and outcome under study. This applies also to social risk factors, so that "it is no longer possible to interpret correlations among biologically related family members as

prima facie evidence of sociocultural causal mechanisms" (Turkheimer 2000, p. 162). Furthermore, researchers who study the influence of biological risk factors (e.g., markers of stress response) or psychological risk factors (e.g., childhood psychological problems) on later outcomes must also account for the possibility that genetic factors could account for the association. We must caution, however, that gene–environment correlations (the *Z* to *X* path in **Figure 1**) do not necessitate that genetic factors confound the statistical association between the risk factors and an outcome. The same genetic factors that influence exposure to the risk factor may not influence or be correlated with the outcome (i.e., there may not be a path from *Z* to *Y*) (Rutter et al. 1993). Nevertheless, gene–environment correlations raise this possibility of genetic confounding.

Confounding by indication: the spurious association between a treatment and an outcome due to the indication (i.e., reason) for the treatment also being a cause of the outcome

For example, there is great interest in the role of prenatal risk factors for later psychological problems (e.g., O'Donnell & Meaney 2017). Yet, quantitative behavior genetic studies have shown that genetic factors influence many prenatal risk factors, such as smoking during pregnancy (D'Onofrio et al. 2003) and birth weight and gestational age (Clausson et al. 2000). Furthermore, genetic factors (as measured by polygenic risk scores, an aggregate of measured genetic influences) associated with neurodevelopmental problems in the mother are also correlated with early-life risk factors for her offspring, including the mother's substance use, use of prescription medications, infections, and stressful life events during pregnancy (Leppert et al. 2019). These findings raise the possibility that common genetic factors could account for associations between prenatal risk factors and later offspring psychopathology (D'Onofrio et al. 2014).

2.5.3. An example of a causal diagram. An example can be found in research exploring the potential consequences of maternal antidepressant use during pregnancy for offspring development (Sujan et al. 2019). We know from existing studies that there is an association between such medication use and offspring autism spectrum disorder. However, this observed association may not reflect a causal effect, as there are several possible noncausal pathways that could explain the association between maternal medication use and offspring neurodevelopment, which we illustrate in **Figure 3**. **Figure 3** first shows that there is a plausible causal effect via putative mediating factors or so-called mechanisms of action (the solid paths in the figure). However, the association could exist for reasons that do not involve a causal effect of maternal medication use (highlighted in the figure by nonsolid paths)—notably, the reason why the women took the medication (the dashed–dotted paths). Epidemiologists refer to such confounding as confounding by indication because the reasons that pregnant women take antidepressants (i.e., maternal anxiety or depression or its causes) also likely affect offspring development independent of medication use. Furthermore, the diagram shows that researchers also must consider confounding from environmental (the dotted paths) and genetic factors (the dashed paths) that are common causes of maternal medication use and offspring neurodevelopment. We note that a more complete causal diagram would entail further specification of the environmental common causes (e.g., poor nutrition). Yet, the diagram illustrates that causal claims about prenatal antidepressant exposure require ruling out a series of alternative explanations (i.e., noncausal pathways), and, notably, this requirement also applies to studies that examine the potential mediating mechanisms.

2.5.4. Measured, unmeasured, and unknown confounding factors. An important principle of causal diagrams is that they must include all known common causes of the putative causal risk factor and outcome regardless of whether a study can measure them. Thus, researchers cannot rely solely on what they are able to measure in a particular study when creating a causal diagram. **Figure 3** highlights the complexity of studying risk factors in observational studies—it is highly unlikely that any single observational study can account for all confounding. The construction of causal diagrams, therefore, can help researchers formally distinguish what a study can and cannot

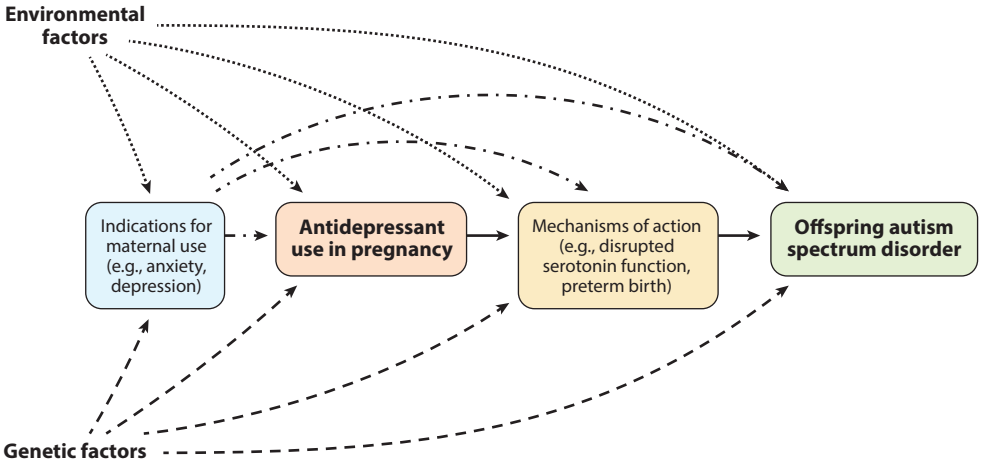

Figure 3

Causal diagram for understanding the processes underlying the association between maternal antidepressant use during pregnancy and offspring autism spectrum disorder. The diagram illustrates how antidepressant use may have a causal effect on offspring autism spectrum disorder via specified mechanisms of action (*solid arrows*). The antidepressant–autism association also could be due to noncausal paths, such as confounding by indication (*dotted–dashed arrows*), environmental factors (*dotted arrows*), and genetic factors (*dashed arrows*). Figure adapted with permission from Sujan et al. (2019).

rule out. Furthermore, there is always the possibility that yet unknown factors cause both the risk factor and the outcome and thus confound the association between the two.

Figure 4 presents a diagram to illustrate these points. Again, the purpose is to estimate the causal effect of X on Y, but the figure replaces the Z (representing all common causes) from **Figure 1** with two sets of variables, C and U, which separate two kinds of variables subsumed by Z. C represents all of the factors that researchers are able to account for in a study (e.g., by adjusting for measured covariates or using design features, as discussed in Section 3). In contrast, U represents all of the common causes that remain—these could be known factors that a study

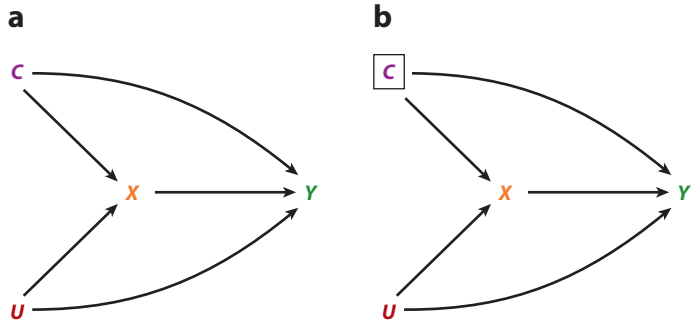

Figure 4

Causal diagram illustrating unmeasured confounding. X represents a risk factor, and Y represents an outcome. C represents all measured common causes of X and Y. U represents all unmeasured common causes. (*a*) The unadjusted association between X and Y would not reflect a causal effect because the variables are also associated through both C and U. (*b*) Conditioning on C, noted by a box, blocks the path from X and Y because of C. However, a conditional association between X and Y (holding constant C) would not reflect the true causal effect because the variables are still associated via U.

cannot address, or they might be unknown factors. Certainly, the unadjusted association between *X* and *Y* in **Figure 4*a*** does not estimate the true causal effect because of the confounding represented by both *C* and *U*. Likewise, a conditional association between *X* and *Y* that accounts for the role of *C* (represented in **Figure 4*b***) does not reflect the true causal effect because of the confounding that remains due to *U*. The amount of bias in estimates of causal effects in conditional models is influenced by the proportion and magnitude of the common causes due to *C* and *U*. Well-designed observational studies seek to carefully select necessary factors to adjust for in order to reduce the influence of unmeasured confounding. However, as illustrated in **Figure 4*b***, observational studies are always limited by the possibility that confounding factors, both known and unknown, could account for the association between *X* and *Y*. This has major implications for the interpretation of observational research.

2.6. Inappropriate Use of Causal Language

Despite calls for a consistent use of terminology (e.g., Kraemer et al. 1997) and cautioning that researchers "are obliged to avoid causal language" (Kendler 2017, p. 562) when presenting the results of observational studies, researchers frequently use language that implies causation. A simple search of published observational studies in clinical psychology, psychiatry, and epidemiology will find a wealth of studies using phrases that imply causation, such as "the effect of" or the "influence of" a risk factor on an outcome. While others may find this language innocuous, we believe the use of such language leads to confusion in the field. And, unfortunately, in some observational studies researchers have drawn explicit causal conclusions even though the studies were unable to account for all (or many) plausible noncausal explanations.

We, as well as other authors (e.g., Cope & Allison 2009, Schwartz et al. 2016), have found two areas of clinical psychology research in which researchers are more likely to use inappropriate casual language or draw unjustified causal conclusions, though certainly not in a majority of the manuscripts. First, research that includes biological functioning, such as measures of physiological functioning, is more likely to make unwarranted causal claims—a phenomenon that is sometimes referred to as neuroseduction (Schwartz et al. 2016). Second, researchers who focus on what might be viewed as righteous goals (e.g., when studying risk factors that may be deleterious for other reasons or viewed as ethically problematic) are also more likely to use causal language when interpreting their findings from observational studies. For example, because the physical and sexual abuse of children is abhorred, readers may not hold claims that such abuse causes an adverse outcome to the same standards for drawing causal inferences. This phenomenon is referred to as white hat bias (Cope & Allison 2009). Again, we believe these are examples of how confirmation bias can lead researchers to fail to acknowledge, account for, and accurately interpret their findings in light of plausible alternative explanations—confounding in particular. In the next section, we review how different analytic approaches and designs can help account for confounding.

3. ANALYTICAL APPROACHES AND DESIGNS TO ADDRESS CONFOUNDING

There are numerous approaches to address confounding that are more or less commonly applied in psychology (Rutter et al. 2001, Thapar & Rutter 2019, West 2009). We consider two broad categories: approaches that directly account for measured confounders and those that seek to indirectly address unmeasured confounding through research design. For each approach, we first provide a description of how it can help account for confounding in observational studies. We then describe the assumptions and limitations and provide an example of how the approach has been used to address a question in clinical psychology.

3.1. Approaches for Accounting for Measured Traits

Researchers can incorporate covariates into their analyses using several approaches to help account for confounding factors. Here we outline three broad approaches within this domain.

3.1.1. Stratified analysis. Although it is not a common practice in psychology, researchers can gain more information about the putative influence of a measured covariate by conducting a stratified analysis, in which researchers restrict their estimation of an association between a risk factor and an outcome to homogenous groups of individuals on the basis of a measured categorical trait. As described above (see Section 2.5.3 and **Figure 3**), the unadjusted association between maternal antidepressant use during pregnancy and offspring functioning (i.e., comparing offspring exposed to the medication with all offspring who were not) may not reflect a causal effect because the association could be due to other factors. To help account for confounding by indication, researchers could stratify or restrict their analyses to the subset of mothers who had diagnoses of depression (this would be the *C* in **Figure 4**). This approach would estimate the association conditional on maternal depression (i.e., the analysis would compare offspring exposed to antidepressants during pregnancy with offspring of women who had depression but did not take the antidepressants during pregnancy). Stratified analyses thus illustrate how researchers can hold constant or condition on a measured variable.

The major limitation of a stratified analysis is that it quickly becomes unmanageable when researchers want to hold constant many factors. The standard way to adjust for more than one covariate and/or continuously measured factors is to use regression models. Broadly, these can be divided into regression models for the outcome and regression models for the exposure.

3.1.2. Outcome regression models. Traditionally, outcome regression models have predominated in the field of psychology (e.g., Cohen et al. 2003). Outcome regression models describe how the outcome depends on the risk factor and the measured confounders (i.e., the arrows from *X* to *Y* and from *C* to *Y*, respectively, in **Figure 4**). To account for confounding due to *C*, outcome regression models block the path from *C* to *Y*. Technically, outcome regression models hold constant the association between the measured covariates *C* and *Y* by partialing out the associations between the covariates and the outcome. Outcome regression models thus provide estimates of the association between a risk factor and an outcome that are adjusted for or conditional on the measured covariates in the model, given certain assumptions (see Section 3.1.5).

3.1.3. Exposure regression models. Exposure regression models are becoming more common in the field. They account for confounding by breaking the path between *C* and *X* in **Figure 4**. These models are almost exclusively used for binary exposures, in which case they describe how the probability of being exposed—the propensity score (PS)—depends on the measured confounders. Using an exposure regression model, the PS can be estimated for all values of the measured confounders, thus giving an estimated PS for each subject in the study.

The popularity of PS methods comes from the fact that even though there may be many measured confounders, the PS is a single variable, and adjusting for this variable completely removes the influence of all measured confounders (Rosenbaum & Rubin 1983). Thus, once the PS has been estimated, researchers may treat it as a single confounder and adjust for it using a standard outcome regression model. Alternatively, researchers might match on the PS. A somewhat different use of the PS is to perform inverse probability weighting (IPW), which artificially breaks the influence of the measured confounders on the exposure by an elaborate weighting scheme. We refer readers to Austin (2011) for an overview of these different PS methods. For example, Brown

et al. (2017) used IPW when studying the association between maternal antidepressant use during pregnancy and offspring autism. The statistical approach accounted for a host of covariates, including medical and psychiatric diagnoses and the use of other prescribed medications during pregnancy, in an attempt to balance the measured covariates among exposed and unexposed offspring. The association with antidepressant use during pregnancy in the conditional model was greatly attenuated compared with the unadjusted model. The results suggested that confounding factors accounted for most of the unadjusted association.

Residual confounding: the systematic error in the estimation of a causal effect due to imperfect adjustment for (i.e., due to measurement error or coarse categorization of) a particular confounder

3.1.4. Outcome regression or exposure regression models? Exposure regression models have advantages for observational research, including the ability to adjust for large numbers of covariates (Lee & Little 2017, West et al. 2014). If both the outcome regression model and the exposure regression model are appropriate, and the sample is large, then these modeling approaches will give similar results. In practice, though, they may give different results due to either (or both) models being misspecified or due to sampling variability in small samples. So which approach is then preferable? In some scenarios, the researcher may know more about the mechanisms behind the exposure than the mechanisms behind the outcome. In practice, outcome regression is more common for three reasons. First, outcome regression easily generalizes to nonbinary (e.g., continuous) exposures, which is not the case for exposure regression. Second, when both regression models are correct, outcome regression models give smaller standard errors. Third, confidence intervals and *P* values are simple to calculate with outcome regression and are provided by all standard statistical software, which is not the case for exposure regression.

3.1.5. Limitations of approaches for accounting for measured traits. While the use of measured covariates is the standard practice within the field of clinical psychology, the use of these statistical approaches is hindered by several key limitations.

3.1.5.1. Inadequate control of common causes. The main limitation of all of the models that adjust for measured covariates, including PS models, is the inability to account for unmeasured confounding (the *U* in **Figure 4**). However, adjusting for measured covariates to account for confounding also is greatly limited for other reasons that many psychologists frequently fail to acknowledge.

First, measured covariates are always measured with error, and this measurement error has serious consequences for testing causal inference using statistical models. Accounting for covariates with measurement error will not account for all of the bias due to the construct. In epidemiology, this is referred to as residual confounding. How deleterious is residual confounding? The answer depends on the quality of measures in each study, but simulation studies suggest that it can be a major concern (Fewell et al. 2007), particularly in the field of psychology (Westfall & Yarkoni 2016). Despite these major reviews, it is not clear that clinical psychology researchers fully appreciate the importance of this limitation. Residual confounding, thus, is a plausible alternative explanation to a causal effect when researchers find conditional associations between a risk factor and an outcome.

Second, researchers often make inappropriate inferences about the degree to which the measured covariates in their study actually represent the higher-order construct they are trying to measure. Whereas residual confounding reflects measurement error, threats to the construct validity of the measured covariates heavily influence the interpretation of regression models. For example, researchers have claimed that statistically adjusting for a trait in parents will account for genetic factors that influence their offspring (in the hopes of accounting for genetic confounding). But there are numerous reasons why this may not be the case, such as findings that the genetic

factors that influence a trait change across the life span (e.g., Hannigan et al. 2017). As such, statistically adjusting for a trait at one point in development does not account for all genetic (or environmental) factors associated with the trait (Silberg et al. 2003).

Finally, fundamental misunderstandings of regression methods have led researchers to make inaccurate claims about the degree to which the methods account for confounding. In particular, prominent researchers have expressed serious concerns about the causal inferences that researchers have drawn based on PS methods (e.g., Luellen et al. 2005, Pearl 2009). Although adjustment for the PS will account for all measured confounders, exposure regression approaches are not a panacea. PS methods make exactly as much (or as little!) adjustment as any of the more traditional, outcome regression methods (Pearl 2009). We note that researchers are debating whether the creation of high-dimensional PSs that incorporate a much wider range of measured covariates than would typically be possible to include as independent predictors (e.g., hundreds of covariates) increases the likelihood that such PSs serve as proxies for relevant unobserved confounders (Schneeweiss et al. 2009).

3.1.5.2. Inappropriate control of measured factors. Although we stress that statistical models that adjust for measured covariates are limited in their ability to account for all confounding, merely adjusting for more measured covariates is not necessarily the solution. Unfortunately, researchers in psychology frequently add numerous covariates to regression models in the hope of adjusting for confounding without understanding the potential harm in doing so (Rohrer 2018). We highlight two important limitations that arise from inappropriate statistical control of measured covariates.

First, researchers should not adjust for mediators (i.e., variables that mediate the effect of X on Y) unless the explicit research question concerns direct effects of the mediator. Mistakenly adjusting for a mediator can bias the estimation of the total effect because part of the causal effect through the mediator has been blocked (Rosenbaum 1984).

Second, adjusting for colliders can lead to bias by inducing a noncausal association between the risk factor and outcome (Greenland et al. 1999). **Figure 5** provides a simplified example of this phenomenon. In **Figure 5*a***, the risk factor X is hypothesized to have a direct causal influence on outcome Y. X also has a causal effect on the variable W. W does not have a causal effect on Y, but the two appear associated because they share the unknown common cause U. The variable W is a collider because it is a common effect of both X and U. Following the causal diagram rules

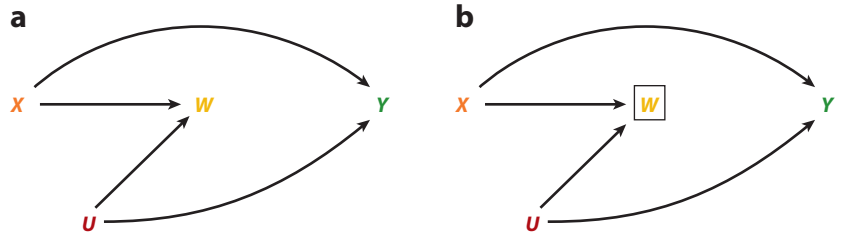

Figure 5

Causal diagram illustrating collider-stratification bias. X represents a risk factor, and Y represents an outcome. W represents an effect of X. U represents unmeasured common causes of W and Y. (*a*) The marginal or unadjusted association between X and Y would represent a causal effect. (*b*) Conditioning on W, noted by a box, opens a noncausal path from X to Y (via W and U). A conditional association between X and Y (holding constant W) would not reflect the causal effect because of the opened path through the collider W.

described above, the path with the collider is closed; hence, the marginal or crude association between *X* and *Y* would represent the causal effect (**Figure 5*a***). However, if a researcher adjusts for the variable *W* (as illustrated in **Figure 5*b***), perhaps to see if it is a mediator, the conditional association between *X* and *Y* would be biased. This is because according to graph theory, conditioning on a collider opens the path through it (*X* would be associated with *Y* via the open path through *W* and *U*). In other words, including variable *W* in an analysis would correlate *X* and *Y* for reasons other than a causal effect. Misestimation of causal effects (i.e., bias) due to controlling for colliders is not frequently mentioned in psychology, and a full exposition is beyond the scope of the current review, but there are several helpful published reviews (Cole et al. 2009, Elwert & Winship 2014). The main point for the purposes of this review is that inappropriately including a measured covariate that is a collider can induce bias in the estimate between a risk factor and an outcome.

Psychologists often assume that adjusting for a measured "third" variable only reduces the statistical association between a risk factor and an outcome (so that the adjusted association would always be a conservative estimate of a causal effect). We stress that this is not always the case. Common causes of a risk factor and outcome can produce both positive and negative associations between the two so that the confounding either inflates or reduces the estimate of a potential causal effect. Moreover, adjusting for a mediator whose effect on the outcome is opposite to the remaining causal effect or adjusting for a collider can also serve to inflate estimates between a risk factor and an outcome.

3.1.6. Review. The use of measured covariates to statistically account for confounding is standard practice in observational research in clinical psychology. We caution, however, that a well-fitting regression model has no relation to whether (*a*) all relevant confounders have been included in the model, (*b*) all relevant confounders have been measured well, or (*c*) all variables included in the model are relevant confounders. Stated differently, a statistical software program cannot differentiate whether a variable is a confounder that is an imprecise measure of an important construct, a confounder that precisely measures an important construct, a mediator, or a collider—the interpretation is based solely on the theory and design of a study. For these reasons, using research designs to control for unmeasured confounders can be a much stronger strategy, especially if such designs are supplemented with analytic approaches to account for measured covariates.

3.2. Designs That Account for Unmeasured Confounders

The important classes of research designs summarized here are often referred to as natural experiments (Rutter et al. 2001, Thapar & Rutter 2019) or quasi-experimental designs (Shadish et al. 2002). These designs frequently use comparison groups that share common causes, regardless of whether they are measured, or examine instances in which exposure to a risk factor is independent of the influence of individuals in the study. As such, these approaches use design features instead of solely relying on measured covariates to account for confounding.

3.2.1. Family-based designs. This powerful class of research design compares relatives within families. Individuals exposed to a risk factor are compared with relatives who were not exposed, which enables these designs to hold constant many factors. Because family members share many genetic variants, family-based designs reduce genetic confounding. In addition, these designs completely control for environmental factors that make the relatives similar. **Figure 6** illustrates a family-based design that includes two individuals (noted by subscript 1 and 2) from a family (noted by subscript *i*) when exploring the association between a risk factor (*X*) and an outcome

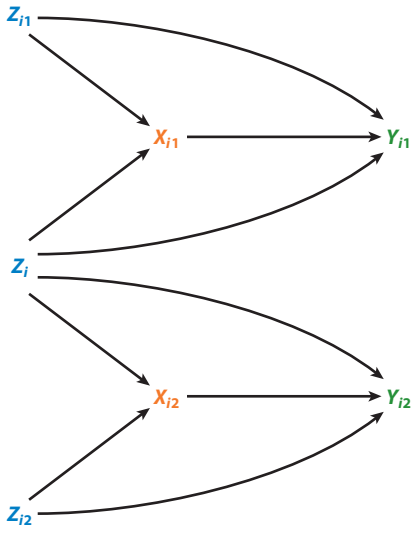

Figure 6

Causal diagram illustrating within-family designs. *X* represents a risk factor, and *Y* represents an outcome, with the subscripts noting that information is available from two individuals (noted by subscript 1 or 2) from the same cluster, in this case a family (noted by subscript *i*). *Z* represents all common causes of both *X* and *Y*. Some of the common causes are factors that are shared by family members (Z_i), and some are not shared by the first or second family members (Z_{i1} and Z_{i2}, respectively).

(*Y*). Family-based designs enable researchers to account for all factors that make the family members similar (noted as Z_i) by design (i.e., the factors do not have to be measured in the study). Comparing family members thus explores the association between a risk factor and an outcome while holding constant all factors that make them similar. The design, however, cannot account for factors that are not shared by the first or second family members (Z_{i1} and Z_{i2}, respectively).

3.2.1.1. Sibling comparisons. Because full siblings share on average 50% of their segregating genes as well as shared environmental factors (i.e., environmental factors that make them similar), the comparison of differentially exposed (i.e., exposure-discordant) siblings enables researchers to account for many common causes that can confound associations between risk factors and outcomes (Lahey & D'Onofrio 2010). For instance, in a study that compared siblings whose mothers used antidepressants during one pregnancy but not during another (while statistically adjusting for measured covariates that were different among the siblings), the association between antidepressant use during pregnancy and offspring neurodevelopmental disorders was found to be small and not statistically significant (Sujan et al. 2017). While the study could not rule out the role of a small causal effect, the finding is generally consistent with other sibling-comparison studies (Brown et al. 2017, Rai et al. 2017) in suggesting that antidepressants do not have a large causal effect on those neurodevelopmental problems; rather, the unadjusted association is due mainly to confounding factors (for a review, see Sujan et al. 2019).

3.2.1.2. Co-twin control design. The comparison of discordant identical twins is a particularly stringent test of causal effects because identical twins share 100% of their genetic makeup as well as environmental factors that make twins similar (McGue et al. 2010). A study by Caspi et al. (2004) provides an illustrative example of how psychologists have used the design, along with other

approaches to rule out confounding (and other possible explanations), when studying the effects of parenting practices on offspring development. Twins who received more negative parenting and less warm parenting had more disruptive behaviors than their identical co-twins, even when the researchers statistically adjusted for measures of previous factors that could account for the association. The use of the co-twin control design and other approaches, therefore, increased the internal validity of the study because the design was able to account for many of the alternative noncausal explanations for the association between parenting and offspring behaviors.

3.2.1.3. Offspring of twins. The offspring of identical twins are genetically related as half-siblings (i.e., they share on average 25% of their segregating genes), though they are socially cousins (D'Onofrio et al. 2003, McAdams et al. 2014). Offspring-of-twins design has been used primarily to study risk factors that are shared by siblings. In a classic study, Gottesman & Bertelsen (1989) compared the offspring of identical twins who were discordant for schizophrenia. Their offspring had the same risk for schizophrenia; this result suggests that the intergenerational transmission of schizophrenia was not due to a causal effect of exposure to parental schizophrenia. In contrast, several offspring-of-twins studies that have explored the intergenerational transmission of depression have found a statistical association that is independent of genetic factors, which is consistent with a causal effect (McAdams et al. 2015, Silberg et al. 2010, Singh et al. 2011).

3.2.1.4. Other family-based designs. There are many family-based designs that are used in the field of behavior genetics to help account for genetic and environmental confounding when studying putative causal risk factors (D'Onofrio et al. 2013, Knopik et al. 2016). For example, adoption studies (e.g., Leve et al. 2013) have enabled researchers to explore the consequences of parental depression on offspring development while holding constant genetic factors due to passive gene–environment correlation (McAdams et al. 2015). Notably, researchers can combine several family-based designs to help account for additional confounding factors and provide information about the extent to which genetic and environmental processes explain the association between a risk factor and an outcome.

3.2.1.5. Limitations of family-based designs. While family-based designs enable researchers to control for potential confounding factors without measuring them, these designs have a number of limitations (D'Onofrio et al. 2013, Frisell et al. 2012, McGue et al. 2010, Sjolander et al. 2016). First, the designs are unable to rule out the influence of factors that are not shared by family members, though researchers can add measured covariates to help account for such individual confounding. These designs are also sensitive to measurement error (or misclassification) and typically require large sample sizes to find enough family members who differ on both the risk factor and the outcome (i.e., they have enough statistical power). Further, they assume no carryover effects from one family member to the other, and they rely on assumptions about whether the findings from the discordant family members generalize to other populations.

3.2.2. Interrupted time series design. The interrupted time series (ITS) design can rule out alternative explanations to allow causal inference by controlling for confounding factors whether they are measured or not (Glass et al. 1975). In an ITS design, an outcome variable is measured repeatedly over time. The time series of such repeated measurements is interrupted if a condition that may causally influence the repeatedly measured outcome changes during the series of measurements. As such, the ITS accounts for confounding factors by breaking the link between *Z* and *X* in **Figure 1**. ITS designs have a mixed history of use in clinical psychology. On the one hand, they played an essential role in the development of behavior therapy (Kazdin & Wilson

1978) and continue to be frequently used to evaluate the efficacy and effectiveness of therapies (Smith 2012). On the other hand, we believe they are greatly underused for studying other risk factors.

The change in condition may be one that is controlled by the experimenter, such as the introduction of a treatment following a series of pretreatment baseline assessments. Much has been written about the advantages of within-person ITS in treatment studies (Kratochwill & Piersel 1983). ITS designs are widely used for this purpose partly because they are often so informative that small numbers of participants can be used in treatment studies, including single-subject designs (Smith 2012). A powerful advantage of their use in treatment outcome research is that the treatment can be introduced at different points in time with different participants to minimize concerns about chance confounding (Kazdin & Kopel 1975).

The interruption also could be an event or condition over which the experimenter has no control, such as an earthquake or a change in laws that regulate the sale of alcohol. For example, during the early 2000s, both Slovenia and Russia implemented laws that substantially restricted the sale of alcohol. Using official records, researchers found that monthly rates of suicide in males during the years after the law changed were markedly lower than in the years before the law changed in both countries (Pridemore et al. 2013, Pridemore & Snowden 2009). Although the event and the change in the outcome can be linked at a particular time and place in such studies, these variables are not linked at the level of the individual. As a result, it is not possible to determine whether the persons with the outcome were the same persons who were affected by the event because not everyone in these countries drank alcohol or experienced changes in their own access to alcohol as a result of the changes in laws.

The fundamental assumption of ITS studies is that no other event occurred at the same time as the event under study that was actually the cause of the change in the outcome. For example, if a change in a law regarding the availability of alcohol happened to occur at a time when employment opportunities were substantially increased in the same country, it would be impossible to determine which was the causal event in that study. In the examples cited in the previous paragraph, the facts that (*a*) the changes in alcohol were associated with the same male-specific decline in suicides in two different countries and (*b*) the laws were changed 3 years apart support the conclusion that decreasing the availability of alcohol saved lives. But the studies do not entirely rule out the possibility that another societal event was the actual cause of the reduction in suicides.

Multiple measurements of the outcome before and after the event are needed to detect possible changes. In the absence of multiple measurements, it is not possible to know whether the difference in the outcome simply reflects changes that were unrelated to the event. There must be a clear and detectable change that occurs at the time of the event to rule out other possible causes of change, such as testing effects, maturation, or regression artifacts. It also is possible to strengthen a causal claim by including a control group in which the event did not occur (e.g., Wing et al. 2018).

3.2.3. Review. Instead of relying solely on measured covariates to account for confounding, family-based designs and ITS studies use design features to help rule out alternative noncausal explanations for an observed association between a risk factor and an outcome. The use of these designs has a long history in the field of psychology. For example, these designs have helped researchers realize that previously identified risk factors did not have causal effects on outcomes (i.e., the associations were due to confounding factors) and have strengthened claims about other causal effects. Unfortunately, we believe that researchers do not use these designs enough in the field of clinical psychology.

4. SUMMARY AND FUTURE DIRECTIONS

Many of the pertinent research questions related to etiology, treatment, and prevention of behavioral health problems cannot be answered by RCTs. In this review, we have sought to introduce advances in the study of causal inference, in particular those related to accounting for confounding in observational studies, to the field of clinical psychology. Below, we emphasize several implications for the field.

4.1. Training in Philosophy of Science

We urge clinical programs to provide more training in aspects of philosophy of science relevant to causal inference (D'Onofrio et al. 2017, O'Donahue 2013), in particular the need to rigorously test competing hypotheses. We believe that one of the primary challenges to causal inference involves the identification of and proper dealing with sources of bias, which require researchers to use analytic and design features to test competing hypotheses in an iterative manner (e.g., Platt 1964).

4.2. Training in and Use of Causal Diagrams

We also have briefly introduced how the use of causal diagrams can help researchers clarify their thinking by formalizing the broader causal theory under which they are working. As such, we encourage more training in causal diagrams and their implementation throughout the field, consistent with other calls in psychology (Rohrer 2018). We propose that the use of causal diagrams would greatly increase the quality of dissertations, manuscripts, and grants focused on observational research.

We frequently hear that a realistic causal diagram would be too complex to be helpful. But we argue that when studying such complex scenarios, it is particularly important to draw a causal diagram for several reasons. First, constructing a causal diagram helps clarify the precise research question. For example, a diagram can help clarify which risk factor or set of risk factors is of primary interest. Second, causal diagrams help researchers identify the status of the field and show if there are any major gaps in the current knowledge. Third, we believe that causal diagrams can help guide the initial design and analysis of studies. For example, causal diagrams can help determine which plausible confounding factors need to be measured (and measured well), which covariates may act as mediators and/or colliders, and what design features would be appropriate to target potential unmeasured confounding. Frequently, the use of causal diagrams highlights the necessity of studying upstream causes of the risk factor to understand the processes through which it may be associated with an outcome. Finally, we have found that causal diagrams foster a level of humility regarding the limitations of studies, which is subsequently reflected in the causal language used (and not used) in the interpretation and conclusion of studies.

We want to provide two examples of how using causal diagrams has influenced our own work. First, identifying the common causes in **Figure 3** greatly aided our research on maternal use of antidepressant medication in pregnancy and offspring neurodevelopmental disorders (Sujan et al. 2017). The figure was also helpful in the review of research on the topic that we were later invited to write (Sujan et al. 2019). Second, the use of causal diagrams has changed how our research team approaches confounder adjustment. When studying the consequences of prenatal risk factors (D'Onofrio et al. 2013), we used to include indicators of maternal and paternal lifetime history of behavioral health problems (in addition to using design features) without realizing that these measures may include behaviors that could have been affected by the risk factor. Such factors require different considerations than confounders (possible mediator and/or collider), and our

consideration of potential confounding factors is now limited to those that occur before the putative causal risk factor.

4.3. Training in and Use of Advanced Statistical and Methodological Approaches to Account for Confounding

We hope that the a priori identification of plausible alternative explanations for observed associations between risk factors and outcomes will help researchers design and analyze data sets using approaches that can better account for theory-driven covariates that are precisely measured. In addition to training in outcome regression models that have generally dominated the field of psychology (Cohen et al. 2003), we encourage more training and use of exposure regression models when appropriate (Lee & Little 2017, West et al. 2014). This will be particularly important as more large-scale data sets with information on behavioral health become available.

Given the fundamental limitations of relying solely on measured covariates, we encourage clinical science researchers to use innovative designs to help account for confounding. This call, which is consistent with previous calls in psychology (e.g., Rohrer 2018, Rutter et al. 2001) and in the social sciences more generally (e.g., Shadish et al. 2002), will require students to receive in-depth training in designs that have a rich history in the field of psychology, such as ITS (Kazdin & Kopel 1975). Clinical psychologists will also need to consider designs that are more common in other fields. For example, behavior genetic researchers use and combine several family-based designs (Knopik et al. 2016). The field of epidemiology also frequently uses negative control exposures, negative control outcomes, and instrumental variables to account for confounding factors (Gage et al. 2016). In addition, research designs used in economics, such as difference-in-difference designs, enable researchers to examine the consequences of broader policies (Wing et al. 2018).

Because each statistical and design-based approach to causal inference has its limitations, researchers will ultimately need to find converging evidence from multiple approaches. If research designs with different threats to their validity reach the same conclusion, the case for causal claims is strengthened. We believe that the causal diagrams will ultimately help researchers from multiple disciplines integrate findings across multiple studies because the diagrams formally present the plausible confounding (and mediating) factors. As such, we encourage researchers to use and combine different methods to account for confounding. For example, researchers have compared the results from a family-based study with those that relied solely on statistical adjustment for covariates (Kendler & Gardner 2010). And the ability to draw strong inferences is greatly aided by research that experimentally manipulates risk factors (e.g., prevention/intervention studies or analog studies), such as studies that compare the results of observational studies with those of RCTs (Ioannidis et al. 2001).

4.4. Training in and Use of Approaches to Account for Other Threats to Validity

Recently, there have been calls in psychology in general and clinical psychology in particular (Tackett et al. 2017) to increase the rigor and replicability of research. We believe that researchers must also acknowledge and address other threats to validity (Shadish et al. 2002). For example, confounding is not the only threat to internal validity. Researchers also need to account for measurement error in the risk factor and the outcome as well as bias due to restricting the sampling strategy or analysis (e.g., conducting a complete case analysis) on a variable that is a collider (Hernan et al. 2004). Researchers also need to consider threats to the statistical validity of studies—the degree to which the association between a risk factor and an outcome has been

appropriately estimated. This includes using appropriate statistical techniques to estimate effect sizes (Cumming 2014) and designing studies with appropriate statistical power (Cohen 1988). Clinical psychology researchers must also carefully consider the construct validity or the degree to which inferences from the operations (e.g., measures, persons, settings) of a study relate to or correspond to higher-order concepts (e.g., Cronbach & Meehl 1955). In addition, researchers must consider threats to external validity or the extent to which inferences about causal effects generalize to other persons, settings, and treatments. This includes trying to ensure, as much as possible, that the participants in a study are representative of the population they are trying to study.

4.5. Role of Clinical Psychology in Epidemiologic Research

Whereas this review focuses on how clinical psychology research could benefit from advances in epidemiology, we also stress that clinical psychologists can greatly enhance the quality of research in epidemiology. Causal inference must be grounded in basic science, and clinical psychologists, with their breadth of training, can guide the development of causal theory with their understanding of basic neuro-, psychological, developmental, and social sciences. Clinical psychologists can also enrich epidemiology research through a better understanding of construct validity. Furthermore, training in translational science can help researchers understand how questions from observational studies can inform and be informed by community-based research to help ameliorate the suffering caused by behavioral health problems. Again, we strongly believe that leveraging advances from multiple disciplines will help advance research in clinical psychology. We hope this review aids in the design, analysis, and interpretation of observational studies that explore the etiology and treatment of clinical problems.

SUMMARY POINTS

Our goal has been to help bridge the gap between research in clinical psychology and epidemiology/biostatistics with a particular emphasis on the critical need for observational research in clinical psychology to rigorously test competing causal hypotheses. Below, we summarize eight key points.

1. Testing and drawing causal inferences requires in-depth content knowledge and theory ahead of time, which can be formally represented and clarified in causal diagrams.
2. Examining causality with observational data requires ruling out alternative hypotheses/explanations of an association between the risk factor and outcome, in particular those that are due to unmeasured and residual confounding.
3. Despite calls for cautious language regarding causal conclusions from observational studies, clinical psychology researchers still frequently imply or claim causal effects that are not justified.
4. Researchers can help account for confounding by using measured variables to model the outcome or the exposure in a regression analysis. Regardless, the reliance on measurement of the confounding factors means the analyses will not account for unmeasured or residual confounding.
5. The inappropriate statistical control of some measured variables (e.g., mediators and colliders) can lead to biased estimates of causal effects. Consequently, statistically adjusting for more variables is not always a conservative approach for estimating associations free of confounders.

6. Because the ability to account for confounding via measured covariates is limited, clinical psychologists need to consider using research designs that account for unmeasured factors.
7. There are several designs to account for different types of unmeasured confounding, and each has its own strengths and limitations.
8. Clinical science researchers have used such designs, along with advanced statistical approaches, to rigorously examine putative causal risk factors by excluding noncausal explanations and identifying the role of confounding factors.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This review was supported by funding from the National Institute on Drug Abuse (R01DA048042), the National Institute of Mental Health (R01MH102221), and the American Foundation for Suicide Prevention.

LITERATURE CITED

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

Bell RQ. 1968. A reinterpretation of the direction of effects in studies of socialization. *Psychol. Rev.* 75:81–95

Brown HK, Ray JG, Wilton AS, Lunsky Y, Gomes T, Vigod SN. 2017. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. *JAMA* 317:1544–52

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. 2015. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47:1236–41

Caspi A, Moffitt TE, Morgan J, Rutter M, Taylor A, et al. 2004. Maternal expressed emotion predicts children's antisocial behavior problems: using monozygotic-twin differences to identify environmental effects on behavioral development. *Dev. Psychol.* 40:149–61

Clausson B, Lichtenstein P, Cnattingius S. 2000. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG* 107:375–81

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum

Cohen J, Cohen P, West SG, Aiken LS. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: Routledge

Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, et al. 2009. Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* 39:417–20

Cope MB, Allison DB. 2009. White hat bias: examples of its presence in obesity research and a call for renewed commitment to faithfulness in research reporting. *Int. J. Obes.* 34:84–88

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52:281–302

Cumming G. 2014. The new statistics: why and how. *Psychol. Sci.* 25:7–29

Deaton A, Cartwright N. 2018. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210:2–21

D'Onofrio BM, Class QA, Lahey BB, Larsson H. 2014. Testing the developmental origins of health and disease hypothesis for psychopathology using family-based, quasi-experimental designs. *Child Dev. Perspect.* 8:151–57

Discussion of the assumptions and limitations inherent in randomized controlled trials.

D'Onofrio BM, Class QA, Rickert ME, Larsson H, Langstrom N, Lichtenstein P. 2013. Preterm birth and mortality and morbidity: a population-based quasi-experimental study. *JAMA Psychiatry* 70:1231–40

Review of how family-based designs can help account for unmeasured confounding factors.

D'Onofrio BM, Lahey BB, Turkheimer E, Lichtenstein P. 2013. The critical need for family-based, quasi-experimental research in integrating genetic and social science research. *Am. J. Public Health* 103:S46–55

D'Onofrio BM, Turkheimer E, Eaves LJ, Corey LA, Berg K, et al. 2003. The role of the Children of Twins design in elucidating causal relations between parent characteristics and child outcomes. *J. Child Psychol. Psychiatry* 44:1130–44

D'Onofrio BM, Viken RJ, Hetrick WP. 2017. Science in clinical psychology. In *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*, ed. MC Makel, JA Plucker, pp. 187–98. Washington, DC: Am. Psychol. Assoc.

Elwert F, Winship C. 2014. Endogenous selection bias: the problem of conditioning on a collider variable. *Annu. Rev. Sociol.* 40:31–53

Fewell Z, Davey Smith G, Sterne JAC. 2007. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am. J. Epidemiol.* 166:646–55

Frisell T, Oberg S, Kuja-Halkola R, Sjolander A. 2012. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology* 23:713–20

Gage SH, Munafò MR, Smith GD. 2016. Causal inference in developmental origins of health and disease (DOHaD) research. *Annu. Rev. Psychol.* 67:567–85

Glass GV, Willson VL, Gottman IM. 1975. *Design and Analysis of Time-Series Experiments*. Boulder, CO: Colorado Assoc. Univ. Press

Gottesman II, Bertelsen A. 1989. Confirming unexpressed genotypes for schizophrenia: risks in the offspring of Fischer's Danish identical and fraternal discordant twins. *Arch. Gen. Psychiatry* 46:867–72

Greenland S, Pearl J, Robins JM. 1999. Causal diagrams for epidemiologic research. *Epidemiology* 10:37–48

Hannigan LJ, Walaker N, Waszczuk MA, McAdams TA, Eley TC. 2017. Aetiological influences on stability and change in emotional and behavioural problems across development: a systematic review. *Psychopathol. Rev.* 4:52–108

Hernan MA, Hernandez-Diaz S, Robins JM. 2004. A structural approach to selection bias. *Epidemiology* 15:615–25

Howards PP, Schisterman EF, Heagerty PJ. 2007. Potential confounding by exposure history and prior outcomes: an example from perinatal epidemiology. *Epidemiology* 18:544–51

Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, et al. 2001. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 286:821–30

Jaffee SR, Price TS. 2012. The implications of genotype–environment correlation for establishing causal processes in psychopathology. *Dev. Psychopathol.* 24:1253–64

Kazdin AE, Kopel SA. 1975. On resolving ambiguities of the multiple-baseline design: problems and recommendations. *Behav. Ther.* 6:601–8

Kazdin AE, Wilson GT. 1978. *Evaluation of Behavior Therapy: Issues, Evidence, and Research Strategies*. Oxford, England: Ballinger

Kendler KS. 2005. Toward a philosophical structure for psychiatry. *Am. J. Psychiatry* 162:433–40

Kendler KS. 2017. Causal inference in psychiatric epidemiology. *JAMA Psychiatry* 74:561–62

Kendler KS. 2019. From many to one to many—the search for causes of psychiatric illness. *JAMA Psychiatry* 76:1085–91

Kendler KS, Gardner CO. 2010. Dependent stressful life events and prior depressive episodes in the prediction of major depression: the problem of causal inference in psychiatric epidemiology. *Arch. Gen. Psychiatry* 67:1120–27

Knopik VS, Neiderhiser JM, DeFries JC, Plomin R. 2016. *Behavioral Genetics*. New York: Worth Publ.

Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. 1997. Coming to terms with the terms of risk. *Arch. Gen. Psychiatry* 54:337–43

Kratochwill TR, Piersel WC. 1983. Time-series research: contributions to empirical clinical practice. *Behav. Assess.* 5:165–76

Lahey BB, D'Onofrio BM. 2010. All in the family: comparing siblings to test causal hypotheses regarding environmental influences on behavior. *Curr. Dir. Psychol. Sci.* 19:319–23

Lee J, Little TD. 2017. A practical guide to propensity score analysis for applied clinical research. *Behav. Res. Ther.* 98:76–90

Leppert B, Havdahl A, Riglin L, Jones HJ, Zheng J, et al. 2019. Association of maternal neurodevelopmental risk alleles with early-life exposures. *JAMA Psychiatry* 76:834–42

Leve LD, Neiderhiser JM, Shaw DS, Ganiban J, Natsuaki MN, Reiss D. 2013. The early growth and development study: a prospective adoption study from birth through middle childhood. *Twin Res. Hum. Genet.* 16:412–23

Luellen JK, Shadish WR, Clark MH. 2005. Propensity scores: an introduction and experimental test. *Eval. Rev.* 29:530–58

Mayo D. 1996. *Error and Growth of Experimental Knowledge*. Chicago: Univ. Chicago Press

McAdams TA, Neiderhiser JM, Rijsdijk FV, Narusyte J, Lichtenstein P, Eley TC. 2014. Accounting for genetic and environmental confounds in associations between parent and child characteristics: a systematic review of children-of-twins studies. *Psychol. Bull.* 140:1138–73

McAdams TA, Rijsdijk FV, Neiderhiser JM, Narusyte J, Shaw DS, et al. 2015. The relationship between parental depressive symptoms and offspring psychopathology: evidence from a children-of-twins study and an adoption study. *Psychol. Med.* 45:2583–94

McGue M, Osler M, Christensen K. 2010. Causal inference and observational research: the utility of twins. *Perspect. Psychol. Sci.* 5:546–56

O'Donahue W. 2013. *Clinical Psychology and the Philosophy of Science*. New York: Springer

O'Donnell KJ, Meaney MJ. 2017. Fetal origins of mental health: the developmental origins of health and disease hypothesis. *Am. J. Psychiatry* 174:319–28

Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. New York: Cambridge Univ. Press

Platt JR. 1964. Strong inference. *Science* 146:347–53

Plomin R, Bergeman CS. 1991. The nature of nurture: genetic influence on "environmental" measures. *Behav. Brain Sci.* 14(3):373–86

Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, et al. 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47:702–9

Popper K. 1962. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books

Pridemore WA, Chamlin MB, Andreev E. 2013. Reduction in male suicide mortality following the 2006 Russian alcohol policy: an interrupted time series analysis. *Am. J. Public Health* 103:2021–26

Pridemore WA, Snowden AJ. 2009. Reduction in suicide mortality following a new national alcohol policy in Slovenia: an interrupted time-series analysis. *Am. J. Public Health* 99:915–20

Rai D, Lee BK, Dalman C, Newschaffer C, Lewis G, Magnusson C. 2017. Antidepressants during pregnancy and autism in offspring: population based cohort study. *BMJ* 358:j2811

Rawlins M. 2008. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet* 372:2152–61

Robins JM. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12:313–20

Rohrer JM. 2018. Thinking clearly about correlations and causation: graphical causal models for observational data. *Adv. Methods Pract. Psychol. Sci.* 1:27–42

Rosenbaum PR. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. A* 147:656–66

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55

Rothman KJ, Greenland S. 2005. Causation and causal inference in epidemiology. *Am. J. Public Health* 95:S144–50

Rutter M. 2000. Psychosocial influences: critiques, findings, and research needs. *Dev. Psychopathol.* 12:375–405

Rutter M, Pickles A, Murray R, Eaves LJ. 2001. Testing hypotheses on specific environmental causal effects on behavior. *Psychol. Bull.* 127:291–324

A comprehensive review of the analysis of causation.

Discussion of how research progress is greatly facilitated by testing and ruling out competing hypotheses.

Introduction to the use of causal diagrams for psychological researchers.

Review of the concepts of causation and tests of causal effects in psychological science.

Rutter M, Silberg J, Simonoff E. 1993. Whither behavior genetics? A developmental psychopathology perspective. In *Nature, Nurture, and Psychology*, ed. R Plomin, GE McClearn, pp. 433–56. Washington, DC: Am. Psychol. Assoc.

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. 2009. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20:512–22

Schwartz SJ, Lilienfeld SO, Meca A, Sauvigné KC. 2016. The role of neuroscience within psychology: a call for inclusiveness over exclusiveness. *Am. Psychol.* 71:52–70

Foundational text on causal inference for social science researchers.

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin

Silberg JL, Maes H, Eaves LJ. 2010. Genetic and environmental influences on the transmission of parental depression to children's depression and conduct disturbance: an extended Children of Twins study. *J. Child Psychol. Psychiatry* 51:734–44

Silberg JL, Parr T, Neale MC, Rutter M, Angold A, Eaves LJ. 2003. Maternal smoking during pregnancy and risk to boys' conduct disturbance: an examination of the causal hypothesis. *Biol. Psychiatry* 53:130–35

Singh AL, D'Onofrio BM, Slutske WS, Turkheimer E, Emery RE, et al. 2011. Parental depression and offspring psychopathology: a Children of Twins study. *Psychol. Med.* 41:1385–95

Sjolander A, Frisell T, Kuja-Halkola R, Oberg S, Zetterqvist J. 2016. Carryover effects in sibling comparison designs. *Epidemiology* 27:852–58

Smith JD. 2012. Single-case experimental designs: a systematic review of published research and current standards. *Psychol. Methods* 17:510–50

Sujan AC, Öberg AS, Quinn PD, D'Onofrio BM. 2019. Annual research review: maternal antidepressant use during pregnancy and offspring neurodevelopmental problems—a critical review and recommendations for future research. *J. Child Psychol. Psychiatry* 60:356–76

Sujan AC, Rickert ME, Öberg AS, Quinn PD, Hernández-Díaz S, et al. 2017. Associations of maternal antidepressant use during the first trimester of pregnancy with preterm birth, small for gestational age, autism spectrum disorder, and attention-deficit/hyperactivity disorder in offspring. *JAMA* 317:1553–62

Review of how propensity scores can help account for measured traits in clinical psychology research.

Tackett JL, Lilienfeld SO, Patrick CJ, Johnson SL, Krueger RF, et al. 2017. It's time to broaden the replicability conversation: thoughts for and from clinical psychological science. *Perspect. Psychol. Sci.* 12:742–56

Thapar A, Rutter M. 2019. Do natural experiments have an important future in the study of mental disorders? *Psychol. Med*. 49:1079–88

Turkheimer E. 2000. Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* 9:160–64

West SG. 2009. Alternatives to randomized experiments. *Curr. Dir. Psychol. Sci.* 18:299–304

West SG, Cham H, Thoemmes F, Renneberg B, Schulze J, Weiler M. 2014. Propensity scores as a basis for equating groups: basic principles and application in clinical treatment outcome research. *J. Consult. Clin. Psychol.* 82:906–19

Review of the limitations of statistically adjusting for measured covariates due to measurement error.

Westfall J, Yarkoni T. 2016. Statistically controlling for confounding constructs is harder than you think. *PLOS ONE* 11:e0152719

Wing C, Simon K, Bello-Gomez RA. 2018. Designing difference in difference studies: best practices for public health policy research. *Annu. Rev. Public Health* 39:453–69