



# Ensuring Causal, Not Casual, Inference

Rashelle J. Musci<sup>1</sup> · Elizabeth Stuart<sup>1</sup>

Published online: 7 January 2019  
© Society for Prevention Research 2018

## Abstract

With innovation in causal inference methods and a rise in non-experimental data availability, a growing number of prevention researchers and advocates are thinking about causal inference. In this commentary, we discuss the current state of science as it relates to causal inference in prevention research, and reflect on key assumptions of these methods. We review challenges associated with the use of causal inference methodology, as well as considerations for hoping to integrate causal inference methods into their research. In short, this commentary addresses the key concepts of causal inference and suggests a greater emphasis on thoughtfully designed studies (to avoid the need for strong and potentially untestable assumptions) combined with analyses of sensitivity to those assumptions.

**Keywords** Causal inference · Randomized controlled trials · Assumptions · Mediation

## Introduction

Many questions in prevention science are fundamentally about causation or causal effects: does a school-based prevention intervention reduce drug use in adulthood? What are the effects of child abuse on children's likelihood of graduating from high school? What are the causal factors that lead to depression? These questions all aim to get at an understanding of what happens under different scenarios, such as how outcomes would differ if an individual receives a prevention program vs. not. Recent advances in causal inference methods and the increasing availability of both randomized controlled trial data as well as intensive longitudinal data have facilitated the use of methods that help researchers get closer to causal effects and not just look at associations between factors; these causal inference methods are becoming increasingly popular among prevention scientists.

Randomized experiments are considered the gold standard for estimating causal effects, as any difference in outcomes can be attributed to the treatment of interest, and not to any pre-existing differences between the treatment and control groups. However, randomization is not always feasible or

ethical and sometimes does not answer the questions of interest. When randomization is infeasible, sophisticated methods (and sometimes strong assumptions) are needed to try to tease apart causal effects. The papers within this special section of *Prevention Science* demonstrate novel areas of causal inference methodology and also indicate several areas that need attention from causal inference methodologists and users of the methods. In this commentary, we provide an overview of causal inference and highlight areas that merit key attention from prevention scientists.

## Current State of the Science

Prevention scientists as a whole are thinking more carefully about causal inference, and this thinking should inform both study design and analytic models. Causal inference tools are particularly useful for prevention scientists as they can help deal with possible confounders when interested in a particular prevention or intervention program or risk or protective factor. Key to causal inference is a clear and precise definition of the effect(s) of interest. In particular, a causal effect is defined as the comparison of outcomes that would be seen if some group of people receives a treatment or intervention of interest (or is exposed to a risk factor of interest) and the outcomes that would be seen *if that same group of people* received (or experienced) the comparison condition instead. These alternative states of the world are known as potential outcomes, with the idea that each person in the population of interest could

---

✉ Rashelle J. Musci  
rmuscil@jhu.edu

<sup>1</sup> Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, 624 North Broadway, Baltimore, MD 21205, USA

receive either treatment condition, and thus either of their potential outcomes could (potentially) be observed (Rubin 1974). This concept of potential outcomes is conceptualized nicely in Bray et al. (2018), whose research question is “What differences in problem alcohol use at age 35 would be expected if all individuals in the population had a certain pattern of reasons for alcohol use at age 19, compared to if all individuals in the population had a different pattern of reasons?” The fundamental challenge of causal inference (Holland 1988) is that each unit (each person at a particular point in time) either receives treatment or control, not both, and thus we observe (at most) one potential outcome for each unit. Causal inference methods, then, aim to estimate causal effects by dealing with this particular type of missing data (the missing potential outcomes). These (fundamentally missing) potential outcomes are also what distinguishes causal inference from traditional statistical inference.

Randomized controlled trials (RCTs) are particularly beneficial for estimating causal effects because they ensure that, at baseline (before the treatment is applied), the treatment and control groups are only randomly different from one another. Thus, any difference in outcomes can be attributed to the treatment itself. Within an RCT standard statistical methods can then be used to compare outcomes across groups and estimate causal effects (with attention to missing data, non-compliance, etc., as needed). Formally, RCTs are useful because they have a known “assignment mechanism:” researchers know that the reason some people ended up receiving the treatment (vs. control) was essentially just a coin flip, unrelated to any outcomes. More nuanced questions can also be asked, including moderation and mediation analyses. Moderation methods help us to understand who interventions work for and under what circumstances, i.e., are there baseline characteristics that are associated with the size of treatment effects. Mediation, on the other hand, tells researchers the mechanisms through which interventions work by examining post-treatment variables (Kraemer et al. 2002). As noted further below, mediation analyses allow the investigation of nuanced questions about how interventions or risk factors operate, but can be challenging even in RCTs because of the strong assumptions required (fundamentally, because the mediator cannot be randomized even if the initial exposure is). This is also true for methods that aim to handle noncompliance in an RCT because, although we can randomize people into one group or another, actually participating in the intervention or program of interest is likely not random (Angrist et al. 1996; Stuart and Jo 2015).

In many cases, randomization is infeasible or unethical (e.g., when studying the effects of childhood maltreatment), and researchers are left relying on non-experimental designs. Causal methods help identify when causal inference is possible, and then try to do so well, including clear articulation of the underlying assumptions. A challenge, though, is that

outside a clean randomized trial with very little noncompliance, nearly any causal inference study will involve some untestable assumptions, which fundamentally relate to the unobserved potential outcomes. For example, propensity score methods rely on an assumption of no unmeasured confounders (Rosenbaum and Rubin 1983), instrumental variables relies on assumptions including that there is no direct effect of the instrument on the outcome (Angrist et al. 1996), and structural equation modeling approaches typically rely on strong parametric modeling assumptions (Vanderweele 2012), which can be especially problematic when exposed and unexposed groups differ on the observed covariates. In some cases, assumptions are testable (e.g., whether two variables have a linear relationship), in which case researchers should take care to test those assumptions. In causal inference, though, key assumptions are often untestable (since they relate to fundamentally unobserved potential outcomes). However, even if untestable, researchers can often consider the plausibility of key assumptions and conduct sensitivity analyses that assess robustness to them (e.g., Liu et al. (2013)). We encourage methodologists to develop more sensitivity analyses to help applied researchers understand the consequences of violations of key causal assumptions.

A key point is that before diving into any particular method or design, it is crucial to clearly state the research question. In particular, we would distinguish between “causal models” (broadly stated) and studies that aim to estimate a causal effect. The latter is often the more policy and practice-oriented and is also a much more tractable problem. It also is the tradition and approach that we (the authors of this commentary) have the most familiarity with, in part, because of a focus on interventions and policies and programs in our own work.

Although there has been some recent attention to a truly causal approach to “causal models” (e.g., Gelman and Imbens 2013), it is generally very challenging to try to estimate a large set of causal associations altogether, or to try to identify “the” cause of a particular outcome. VanderWeele (2012) provides a thorough discussion of this, including all of the embedded assumptions when trying to estimate a large “causal model.” In practice, it is often more tractable to try to estimate a particular causal effect—isolating the difference in outcomes associated with a particular risk factor or exposure or treatment. Given these challenges, we focus on estimating causal effects in this commentary. And in fact, as noted by Gelman and Imbens (2013), many “reverse causal questions” about whether Y was caused by X can be turned around into (more tractable) questions of whether X (vs. X’) leads to different levels of Y. There is value in more exploratory analyses that help identify the potential intervention points or factors to focus on for intervention, and many of the causal modeling approaches have that potential value, to identify potential factors to intervene on, which can then be followed up in subsequent studies focused on estimating causal effects.

The papers in this special section address some causal inference problems encountered by researchers in prevention science. In this commentary, we aim to both look back at where prevention science is in terms of causal inference and also provide a path forward for ensuring appropriate and accurate causal inferences moving forward.

## Overview of Special Section

This special section contains a collection of papers that address a number of advances in causal inference methodology and how these advances may relate to prevention science. The papers each focus on important and trending topic areas, including latent variable methods, machine learning, mediation, intensive longitudinal data, and power calculations.

A common challenge in prevention science is latent variables: situations where we cannot directly observe some variable of interest (e.g., depression, aggressive behavior), and instead, observe only proxies (or a scale) related to that underlying latent construct. In their paper, Bray et al. (2018) tackle the scenario where the exposure itself is latent, building on previous work by Schuler et al. (2014), among others. This work, thus, has implications for any studies in prevention science interested in estimating the causal effects of latent factors; this could include studying the effects of childhood violence exposure or the effects of depression in adolescence on adult outcomes. Utilizing a nationally representative sample of high school students, Bray et al. (2018) use inverse propensity score weighting to account for confounding among the complex relationship between latent class membership and a distal outcome of interest. Some causal inference methodology, such as propensity score methods, work to equate treatment and control groups, but this becomes complicated when the treatment or exposure is latent, as it is in this case. This method, and scenario, also raises interesting philosophical questions and debate, as some in the causal inference world would argue that the exposure of interest needs to be a well-defined intervention that could be given or withheld from each individual. Interpretation of effects may then be challenging in these sorts of situations (Hernán and Taubman 2008), although it could still be worth comparing groups who are as similar as possible on the observed characteristics, even if differences in outcome are not interpreted as “causal” per se. The key next steps to extend this work could lie in sensitivity analyses to explore unobserved confounding for this setting with latent treatments.

Power analyses are one of the least liked parts of research for most scientists, but they remain an important aspect to planning and executing data collection and data analysis. Kelcey et al. (2018) demonstrate methods for sample size planning when the goal is to explore multilevel mediation relationships. The methods have relevance for studies

interested in multilevel mediation, such as examining whether the effects of a classroom behavior intervention on children’s outcomes are mediated by the teachers’ behavior. With increasing emphasis on examining mechanisms, particularly from key funding agencies (e.g., National Institute of Mental Health 2018), researchers must pay careful attention to adequately powering studies for mediation and moderation effects. Kelcey et al. (2018) use commonly used power calculation packages (e.g., PowerUp software) to demonstrate estimation of power within a cluster randomized intervention. A challenge with this type of analysis is that while the intervention or prevention is randomized, the mediator is not and therefore, causal inference is challenging. Future work should focus on greater exploration of intraclass correlations of the mediator as well as extending these power analyses to newer mediation approaches, which define causal estimands more clearly and relax some of the assumptions.

Wiedermann and colleagues (Wiedermann et al. 2018) explore mediation when the treatment is randomized, and the mediator and outcome are measured cross-sectionally. Applications could include many situations where a randomized trial is conducted but with just one follow-up time point when both potential mediators and outcomes are measured at the same time. This study design is challenging given the key assumptions about temporality in causal inference methods; without clear temporal ordering, we do not know if the mediator causes the outcome or vice versa. Without a design that ensures temporality, strong modeling assumptions must be made (some of which are examined by Wiedermann et al.). Future work should explore the robustness of results to those assumptions, such as the exact model forms (such as linearity) and the distribution of error terms. Because of these challenges and underlying assumptions, a design that ensures temporal ordering of exposure, mediator(s), and outcome(s) should be preferred, or at least the question wording done in such a way that the mediator measurement refers back in time (e.g., asking individuals about their behavior in the past 30 days and the outcome measured on the actual measurement day). For example, a study exploring whether post-partum depression mediates the relationship between early interventions and child development outcomes would ensure temporal ordering if the study is designed to capture post-partum depression during the post-partum period rather than asking the mother to recall their symptoms.

Combining causal inference methodology with machine learning results in causal structure learning, which may be particularly useful for researchers in prevention science where causal mechanisms are a common research question of interest. This approach may be of interest when researchers aim to understand causal relationships between a large number of variables, such as depression, anxiety, and suicidal behaviors among adolescents. Shimizu (2018) utilizes causal structure learning by demonstrating the use of the linear non-Gaussian

acyclic model, called LiNGAM. The goal of these analytic models is to estimate causal structures of variables in the presence of unobserved common causes. They are typically used in hypothesis generation rather than hypothesis testing. These complex models are based within the structural equation framework and therefore, careful consideration must be taken to ensure that the research question at hand is useful (Vanderweele 2012). The methodology presented by Shimizu (2018) identifies only the class of models that are identified by the data and assumptions, but there may still be concerns about the parametric and model assumptions required to even identify that class.

With increasing popularity of intensive longitudinal data collection tools like ecological momentary assessment, prevention scientists are looking for analytic tools to appropriately model this complex data. Molenaar (2018) uses vector autoregressive (VAR) modeling within a Granger causality testing framework for such data. Granger causality suggests that relationships between variables are autoregressive and cross-lagged in nature such that if a variable  $X_1$  is related to  $X_2$ , then past values of  $X_1$  should predict  $X_2$  above and beyond past values of  $X_2$  alone (Ding et al. 2006). The empirical example discussed by Molenaar (2018) demonstrates Granger causal influence on a child's electro-dermal activity (EDA) within a setting of occupational therapy. With increasing availability of intensive longitudinal data (e.g., combining data on social media posts with data on where adolescents are geographically, or understanding how mood relates to cigarette smoking in a complex longitudinal setting), prevention methodologists will need novel methods to analyze such complex data.

## Moving the Field Forward

There are a number of strategies that prevention science should consider moving forward to enhance the ability of the field to draw appropriate and accurate causal inferences. It is crucial that prevention scientists do two things whenever in the realm of causal inference. First, prevention scientists should first clearly articulate their research question. One framework for doing so, which is highly consistent with years of writings by Donald Rubin and Paul Rosenbaum, is to “emulate a randomized trial” (Hernán and Robins 2016). When designing and implementing a randomized trial, researchers are very familiar with the need to very clearly state the intervention and comparison conditions, the population of interest, and the outcome(s) of interest. These steps are just as (or more) important in non-experimental studies, but are often swept aside as people simply run “causal models” on their data, with little attention to the causal effects of actual interest and how to isolate those effects.

Second, prevention scientists need to understand the assumptions underlying any approach, and there need to be diagnostics to assess those assumptions (when possible), and sensitivity analyses to assess robustness to the assumptions. When using propensity scores, for example, covariate balance diagnostics can be used to assess whether the propensity score approach “worked” at creating treatment and comparison groups that are similar on the observed covariates (e.g., Austin and Stuart 2015; Stuart 2010). And then, importantly, sensitivity analyses can be conducted to assess how sensitive the results are to a potential unobserved confounder (e.g., Rosenbaum 2005; VanderWeele and Ding 2017). Methodologists developing new causal methods should think carefully about appropriate sensitivity analyses for their approach, and then these sorts of diagnostic and sensitivity analysis tools need to be made accessible for prevention scientists, and then used by them.

To help facilitate researchers' ability to match study designs to research questions and understand the underlying assumptions, it is important that researchers know about the spectrum of designs that allow estimation of causal effects and their underlying assumptions. This will ensure that researchers can map the best design onto the research question. Nice overviews of designs can be found in West et al. (2008), Imbens and Rubin (2015), and Rosenbaum (2017). It is not enough for students to learn about RCTs and SEM; they need to be able to identify scenarios where instrumental variables may be most appropriate, when a regression discontinuity design may work, and how to design a high-quality comparative interrupted time series design.

One way of being able to pick the right design for the study, and to assess the plausibility of the assumptions, is to have deep scientific knowledge about the subject area. We thus encourage close collaborations between methodologists and subject matter experts. Causal inference is hard, relying on untestable assumptions about fundamentally unobservable potential outcomes, and you cannot just throw data at the methods—you need to be able to assess the validity of those assumptions using subject matter knowledge.

While prevention science researchers are known for their scientific rigor and reproducibility, a greater emphasis should be placed on transparency in analytic tools. This could easily be accomplished by making analytic code or software easily available, as many of the authors in this special section did. Availability of code for complex models will ensure that the methods and tools get used, and hopefully used appropriately. Further, a common thread through most of this commentary is the need for careful thinking surrounding assumptions of given analytic method, which is particularly important for non-experimental studies and mediation analyses. Researchers must be clear about the assumptions and develop or use sensitivity analyses to assess sensitivity to the key assumptions.

There are also clear lessons for study design. Powering the study for mediation, and not just for exploring main effects, is essential now that there is a greater emphasis on understanding mechanisms. Measuring confounders, including of the mediator/outcome relationship, is essential, even in an RCT. While covariates are not technically needed in an RCT for unbiased effect estimates of the overall treatment effect, they are crucial for answering more nuanced questions like mediation. At the end of the day, it is clear that a well-designed study “trumps analysis” when it comes to objective causal inference and therefore, prevention scientists should aim for thoughtfully designed studies that can be analyzed with a clear focus on the assumptions of the analytic method (Rubin 2008).

**Acknowledgements** The authors thank Wolfgang Wiedermann for the invitation to submit this commentary.

**Funding** Dr. Stuart’s work on this commentary was supported by the National Institute of Mental Health, R01MH115487 (PI: Stuart).

## Compliance with Ethical Standards

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Ethical Approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed Consent** Because this article is a commentary, informed consent is not applicable.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*, 3661–3679. <https://doi.org/10.1002/sim.6607>.
- Bray, B.C., Dziak, J.J., Oatrick, M.E., and Lanza, S. T. (2018). Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. *Prevention Science*.
- Ding, M., Chen, Y., & Bressler, S. L. (2006). Granger causality: Basic theory and application to neuroscience. In *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications* (pp. 437–460).
- Gelman, A., & Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions*. Retrieved from <http://www.nber.org/papers/w19614.pdf>.
- Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*, 758–764. <https://doi.org/10.1093/aje/kwv254>.
- Hernán, M. A., & Taubman, S. L. (2008). Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, *32*, S8–S14. <https://doi.org/10.1038/ijo.2008.82>.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology* (volume 18, pp. 449–484). American Sociological Association.
- Imbens, G.W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Kelcey, B., Spybrook, J., & Dong, N. (2018). Sample size planning for cluster-randomized interventions probing multilevel mediation. *Prevention Science*.
- Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–883.
- Liu, W., Kuramoto, S. K., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prevention Science*, *14*, 570–580 <http://www.ncbi.nlm.nih.gov/pubmed/23408282>.
- Molenaar, P. C. (2018). Granger causality testing with intensive longitudinal data. *Prevention Science*, 1–10. <https://doi.org/10.1007/s1121-018-0919-0>.
- National Institute of Mental Health (2018). *Clinical trials to test the effectiveness of treatment, preventive, and services interventions* (R01 clinical trial required). Retrieved July 20, 2018, from <https://grants.nih.gov/grants/guide/rfa-files/RFA-MH-18-701.html>.
- Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In *Encyclopedia of Statistics in Behavioral Science* (pp. 1809–1814).
- Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*, 808–840.
- Schuler, M. S., Leoutsakos, J. M. S., & Stuart, E. A. (2014). Addressing confounding when estimating the effects of latent classes on a distal outcome. *Health Services and Outcomes Research Methodology*, *14*, 232–254. <https://doi.org/10.1007/s10742-014-0122-0>.
- Shimizu, S. (2018). Non-Gaussian methods for causal structure learning. *Prevention Science*, 1–11. <https://doi.org/10.1007/s1121-018-0901-x>.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21. <https://doi.org/10.1214/09-STS313>.
- Stuart, E. A., & Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, *24*, 657–674. <https://doi.org/10.1177/0962280211421840>.
- Vanderweele, T. J. (2012). Invited commentary: Structural equation models and epidemiologic analysis. *American Journal of Epidemiology*, *176*, 608–612. <https://doi.org/10.1093/aje/kws213>.
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*, 268. <https://doi.org/10.7326/M16-2607>.
- West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., ... Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*(8), 1359–1366. <https://doi.org/10.2105/AJPH.2007.124446>
- Wiedermann, W., Li, X., & von Eye, A. (2018). Testing the causal direction of mediation effects in randomized intervention studies. *Prevention Science*, 1–12. <https://doi.org/10.1007/s1121-018-0900-y>