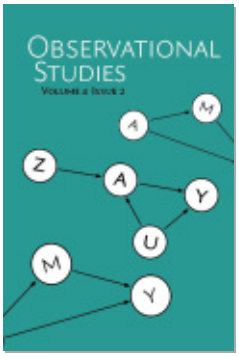
The Tenability of Counterhypotheses: A comment on Bross' discussion of statistical criticism



Jennifer Hill, Katherine J. Hoggatt

Observational Studies, Volume 4, Issue 2, 2018, pp. 34-41 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2018.0004*

Observational Studies 4 (2018) 34-41 Submitted 1/18; Published 1/18

The Tenability of Counterhypotheses: A comment on Bross' discussion of statistical criticism

Jennifer Hill **jennifer.hill@nyu.edu**
Department of Applied Statistics, Social Science, and the Humanities
New York University
New York, New York, 10003

Katherine J. Hoggatt **katherine.hoggatt@va.gov**
VA Health Services Research and Development (HSR&D)
Center for the Study of Healthcare Innovation, Implementation & Policy
VA Greater Los Angeles Healthcare System
Los Angeles, CA, 90073

1. Introduction

We enjoyed reading the commentary by Bross about the appropriate role for a statistician as critic (Bross, 1960). It was an important discussion to initiate at the time the article was written, particularly in light of the highly contentious and scientifically critical debate about the link between smoking and cancer, which involved some of the leading statistical minds of the era (Cornfield, 1954; Cornfield et al., 1959). We believe discussions about the role of a statistical critic are equally if not more relevant today given that our ability to casually critique the work of research "proponents," to use Bross' term, and to disseminate such comments broadly in unrefereed venues has increased exponentially since the time that Bross was writing. Given the complexity and breadth of the issues involved, we focus our discussion on Brosss contention that a critic should have a tenable counter-hypothesis. We further position our comments within the context of causal inference where some additional subtleties arise with regard to satisfying this requirement.

2. Tenable counter-hypotheses

For a critic's counter-hypothesis to be tenable, Bross maintains that "a minimal requirement would be that the effects predicted from the critic's hypothesis should be in line with the actual data, at least in direction and order of magnitude." This seems reasonable in the abstract, but there is not a well-defined criterion for meeting this requirement. For example, how should we achieve this goal if we wish to estimate a causal effect from observational data when even reliably estimating the direction of effects requires making strong, often untestable, assumptions? To illustrate these issues, we consider the common scenario where a proponent is investigating a hypothesis about a causal effect of an exposure on an outcome using observational data. We describe how "dogmatic" statistical criticism (e.g., that you cannot infer causation from observational data) can lead to further methodological errors

and fail to shed light on whether a proponent's hypothesis or a critic's counter-hypothesis should be considered more credible. Finally, we discuss how sensitivity analysis may provide a path forward.

2.1 Observational Data

Let us consider a specific situation in which the proponent's initial hypothesis is about a causal effect. For example, suppose that a proponent claims that exposure to the measles, mumps, and rubella (MMR) vaccine increases the occurrence of autism. Let us also assume for the moment that no randomized or natural experiment is available. According to Bross' criterion, a critic who disagreed with the proponent's claim would need to posit a counter-hypothesis, such as that the MMR vaccine has no effect on autism incidence. Further, according to Bross, this counter-hypothesis should be supported by, or at least not inconsistent with, existing observational data. Let us assume that the critic has access to a reasonably-sized, child-level observational dataset with accurate measurements of what vaccines the child received (and when), subsequent developmental assessment resulting in an autism diagnosis, and pre-treatment measurements of potential confounders. What kinds of estimates from this dataset might we be willing to accept as supportive of a counter-hypothesis?

Even ignoring the (not insubstantial) issues around statistical and practical significance (Berger and Selke, 1987; Gelman and Loken, 2013; Wasserstein and Lazar, 2016), major issues loom. We know that any given estimate of $\mathrm{E}[Y \mid Z, X]$ (where Y is the outcome, Z is the treatment, and X is a vector of potential confounding covariates unaffected by the treatment) is unlikely to be an unbiased estimate of the true estimand, e.g. $\mathrm{E}[Y(1) - Y(0)]$ (where $Y(0)$ and $Y(1)$ are potential outcomes with the typical definitions, as in Rubin, 1978). There are several reasons for this, but first and foremost it is unlikely that we have satisfied the so-called ignorability assumption, $Y(1), Y(0) \perp Z|X$. Colloquially speaking (and ignoring some technical subtleties, see, for instance Greenland et al., 1999), this means it is unlikely in most observational studies that we have measured all confounding covariates.

In the absence of a design that creates this independence structure, we are left to consider how estimates of the causal estimand behave when we only control for subsets of X that are insufficient to guarantee ignorability. For example, suppose the truth is that vaccines decrease the incidence of autism (even though the marginal association is positive). This could lead to a situation where analyses that include a proper subset of the sufficient set confounders yield estimates that are not only biased but of the wrong sign. Such a situation (which arguably is not terribly rare) would make it all too easy to use the data as "evidence" that supports a variety of different counter-hypotheses; that is, it would be easy to show that the estimands corresponding to a counter-hypothesis of a positive effect are "in line with the data."

A more confusing situation arises when we posit a point hypothesis (as opposed to the directional hypothesis above), such as that the true effect of MMR vaccine on autism is 0 (for a discussion of the evidence that supports this claim see Plotkin et al., 2009). Supporting such a counter-hypothesis would be complicated, not only because estimates from analyses that do not satisfy ignorability might have different signs and magnitudes, but also because it is unclear what it means to support a point null hypothesis.

2.2 Randomization to the rescue?

How can we proceed if it is uncertain or unlikely that ignorability is satisfied? An overly simplistic solution to the problem might be to require that only evidence from randomized experiments be accepted to support a counter-hypothesis. After all, in its pure form the randomized experiment justifies the assumption of (strong) ignorability. Bross (1960) and other contemporaries (including, notably, (Cornfield, 1954)) expressed frustration however that statisticians were using the "gold standard" of the randomized experiment as a cudgel to beat down all attempts to make a causal claim using observational data. In fact Bross (1960) highlights this practice in the article in the section on "dogmatic criticism".

We are concerned by reflexive dogmatic criticism as well. One problem with the emphasis on a controlled or natural experiment is that we may ignore evidence about causal effects if that evidence is derived from non-randomized experiments. This tunnel vision can be particularly problematic when investigating research questions that do not lend themselves to randomized experiments for ethical or logistical reasons. An additional problem is that we may end up overstating the infallibility of randomized (controlled or naturally occuring) experiments that occur in practice, no matter their vulnerabilities. Many complications can and often do arise that would preclude a researcher from making a causal claim, even in the context of a randomized experiment, without making additional assumptions. These complications include but are not limited to missing data, noncompliance, measurement error, and grouped data structures. Combinations of these issues are common and are even more difficult to handle (for example, see Barnard et al., 2003; Reardon and Raudenbush, 2013). In situations where the randomized experiment is free from such complications or when additional required assumptions seem plausible (e.g. the exclusion restriction in a randomized experiment with noncompliance or a missing at random assumption to recover missing data), randomized experiments are nonetheless almost always limited in their generalizability (Stuart et al., 2015).

Even given these well-known limitations, there persists a belief that a study with randomization is necessarily a more rigorous approach to a causal inquiry than a study without this feature (Imai et al., 2008). This confidence tends to extend to so-called natural experiments as well (see, for example, Duncan et al., 2004), including methods such as instrumental variables, regression discontinuity, and even, oddly, fixed effects models for identifying causal effects. Yet we know that when the assumptions of these methods fail to hold, things can go badly quickly (Angrist et al., 1996; Reardon and Raudenbush, 2013; Martens et al., 2006; Middleton et al., 2016). Moreover, even when one of these methodologies works well, it will, like randomized experiments, tend to yield estimates that apply most directly to narrow slices of the observation sample, and additional assumptions are necessary to generalize these local average treatment effect (LATE) estimates to a broader population (Hoggatt and Greenland, 2014).

The upshot is that when making causal inferences, most analyses, whether they use data from observational studies or randomized experiments, will rely on some sort of untestable assumptions. If we are requiring that a counter-hypothesis be tenable, it seems the criteria should include a reasonable assessment of the plausibility of such assumptions. However, if we are comparing competing sets of untestable assumptions (corresponding to the proponent's original analysis and the critic's analysis in support of a counter-hypothesis) how

should we assess which of the sets of assumptions are most plausible? Would it be better, for instance, to use an instrumental variables approach where the instrument is weak and the exclusion restriction is questionable or to use an observational study where we are uncertain that we have measured all confounders?

2.3 Sensitivity Analysis: A way forward?

One way to tackle this problem is to promote increased use of sensitivity analyses, by which we mean any of a variety of approaches that explore the sensitivity of our estimates to violations of key assumptions of our analysis. Rather than making a binary decision about which counter-hypotheses are tenable, the goal would be for critics (and perhaps proponents if they are acting as their own critics) to provide a range of estimates that are derived from different sets of assumptions supporting the proponent's analyses. Bross was one of the first scholars to propose this strategy in the context of health research (Bross, 1966, 1967), and much of the early sensitivity analysis literature focused on methods to address possible departures from ignorability assumption (for example Cornfield et al., 1959).

Yet today, more than 50 years after Bross wrote his commentary, sensitivity analysis is seldom used in applied empirical research. This is true despite that the fact that there has been increased focus in the methodological literature in recent years on approaches to assess sensitivity to departures from the ignorability assumption in simple observational studies (see, for example, Rosenbaum and Rubin, 983a; Rosenbaum, 1987; Greenland, 1996; Gastwirth et al., 1998; Rosenbaum, 2002; Imbens, 2003; McCandless et al., 2007; Rosenbaum, 2010; Harada, 2013; Carnegie et al., 2016; Dorie et al., 2016). Moreover some simple sensitivity analyses can be done with a basic spreadsheet program, and software packages are available for more complex applications (for example, Gangl, 2004; Keele, 2010; Carnegie et al., 2015).

It is true, however, that there has been less of a focus on developing methods and software to explore the sensitivity to assumptions required for other types of causal analyses including instrumental variables, mediation, fixed effects, and regression discontinuity (exceptions include Imbens and Rubin, 1997; Small, 2007; Imai et al., 2010; Middleton et al., 2016; McCandless and Somers, 2017). Even more rare are publications that compare two competing identification strategies (for an interesting example of this see DiPrete and Gangl, 2004) or that simultaneously address two different types of assumptions in one analysis (for example Dorie et al., 2016). Certainly more work is needed to create user-friendly, interpretable approaches that can be applied in a variety of circumstances.

Furthermore, the results from a sensitivity analysis will be more useful when that analysis incorporates both statistical expertise and a subject matter expert's prior knowledge A truly interdisciplinary approach to sensitivity analysis could, for example, incorporate subject-matter-specific models for the data generating process, information about the types of likely unobserved confounders, and the most plausible direction and magnitude of (conditional) associations between the unobserved confounders and the treatment and outcome. It could also address Bross' call for the statistical critic to supply more specific and tenable counter-hypotheses. Unfortunately, requiring collaboration between investigators, who are subject matter experts, and statisticians, who can translate this knowledge into parameters

for a formal, quantitative sensitivity analysis, may create an additional hurdle to broader use.

We argue that the practical barriers to adoption of sensitivity analysis are not merely technical, however. Equally lacking are clear incentives to make sensitivity analysis a routine part of empirical research. It is understandable that research proponents would be reluctant to incorporate additional analyses that may make their findings less credible. For example, "failure testing" to assess quantitatively whether a violation of ignorability could "explain away" an observed association will often show that the existence of such a confounder is possible, if not plausible. More sophisticated applications of sensitivity analysis may be better suited to the objectives of research proponents when these methods quantify how causal effect estimates change depending on a wide range of specific assumptions (for example see Carnegie et al., 2016), thus formalizing the process of assessing which counter-hypotheses are most tenable.

External incentives to promote sensitivity analysis may also be needed, and research gatekeepers (such as editors and reviewers) have an important role to play. For example, editorial standards could promote the use and transparent reporting of results from sensitivity analyses in the peer-reviewed literature. Referees of papers could request statistical reviews and encourage or even require that research proponents include sensitivity analysis for key assumptions. Journal editors could also require that statistical criticism be published in discussion articles with clearly stated counter-hypotheses and sensitivity analysis as appropriate. Efforts to encourage data sharing can also promote more rigorous evaluation of counter-hypotheses using a research proponent's own data.

3. Conclusion

Today, as when Bross wrote his commentary, it can seem as if the only job of the statistical critic is to point out problems that could occur, regardless of plausibility or likely impact. A downside of this kind of hit and run criticism, where the mere observation of a flaw in a study's methodology is enough to discount the study's findings, is that it can foster a double standard whereby a research proponent must rule out every conceivable alternative hypothesis to justify a study's findings but a critic need only suggest a counter-hypothesis to undermine them. This double standard may lead to knee-jerk dismissal of findings based on observational data and overconfidence in randomization as a design feature. We agree with Bross that progress may require the statistical critic to stand on more equal footing with the research proponent. Incentivizing or requiring this has ramifications for issues as broad as the standards for peer review in journal publication, data sharing policies, and establishment of criteria for evaluating the empirical support for scientific hypotheses.

Acknowledgments

We would like to acknowledge support for this project from VA HSR&D (IIR 15-436) as well as by Institute of Education Sciences grant R305D110037. The views expressed within are solely those of the authors, and do not necessarily represent the views of the Department of Veterans Affairs or of the United States government.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–472.

Barnard, J., Frangakis, C., Hill, J. L., and Rubin, D. B. (2003). A principal stratification approach to broken randomized experiments: A case study of vouchers in new york city. *Journal of the American Statistical Association*, 98:299–323.

Berger, J. and Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82:112–122.

Bross, I. D. (1960). Statistical criticism. *Cancer*, 13:394400.

Bross, I. D. (1966). Spurious effects from an extraneous variable. *Journal of Chronic Diseases*, 19(6):637–647.

Bross, I. D. (1967). Pertinency of an extraneous variable. *Journal of Chronic Diseases*, 20(7):487–495.

Carnegie, N. B., Harada, M., Dorie, V., and Hill, J. (2015). *treatSens: Sensitivity Analysis for Causal Inference*. R package version 2.0, accessed 07/13/2015. Available from: http://CRAN.R-project.org/package=treatSens.

Carnegie, N. B., Harada, M., and Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9:395–420.

Cornfield, J. (1954). Questions and answers: Statistical relationships and proof in medicine. *The American Statistician*, 22:173–203.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203.

DiPrete, T. and Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34:271–310.

Dorie, V., Carnegie, N. B., Harada, M., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics and Medicine*, 35:3453–3470.

Duncan, G. J., Magnuson, K. A., and Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1(1-2):59–80. Available from: https://doi.org/10.1080/15427609.2004.9683330.

Gangl, M. (2004). Rbounds: Stata module to perform rosenbaum sensitivity analysis for average treatment effects on the treated. Available from: https://EconPapers.repec.org/RePEc:boc:bocode:s438301.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. Technical report, Columbia University.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25(6):1107–1116.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Harada, M. (2013). Generalized sensitivity analysis. Technical report, New York University, New York, NY.

Hoggatt, K. J. and Greenland, S. (2014). Extending organizational schema for causal effects (commentary to accompany gatto, campbell, and schwartz). *Epidemiology*, 25(1):98–102.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2):481–502.

Imbens, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. In *The American Economic Review: Papers and Proceedings of the One Hundred Fifteenth Annual Meeting of the American Economic Association*, volume 93, pages 126–132, New York, NY. American Economic Association.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25:305–327.

Keele, L. (2010). An overview of rbounds: An r package for rosenbaum bounds sensitivity analysis with matched data. Technical report, Columbus, OH.

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. (2006). Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11):2331–2347.

McCandless, L. C. and Somers, J. (2017). Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. *Statistical Methods in Medical Research*, in press.

Middleton, J., Scott, M., Diakow, R., and Hill, J. (2016). Bias amplification and bias unmasking. *Political Analysis*, 24:307–323.

Plotkin, S., Gerber, J., and Offit, P. (2009). Vaccines and autism: A tale of shifting hypotheses. *Clinical Infectious Diseases*, 48(4):456–461.

Reardon, S. and Raudenbush, S. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods and Research*, 42(2):143–163.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York.

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105:692–702.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6:34–58.

Small, D. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102:1049–1058.

Stuart, E., Bradshaw, C. P., and Leaf, P. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16:475–485.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician*, 70:129–133.