The Challenge of Designing Quality Guidelines for Observational Research

John F. Pfaff[*]

As the evidence based policy movement has grown, so too has the volume of quality guidelines produced for randomized clinical trials (RCTs): the National Guideline Clearinghouse alone provides access to over two *thousand* of them. Yet almost no work has been done to design quality guidelines for non-experimental, observational research.[1] This is a troubling (non-) development, since in many ways observational work requires quality guidelines more than experimental. In this paper I will briefly explain why it is essential to create observational quality guidelines, and I will lay out the substantial challenges that such a task faces—and argue that these very difficulties only emphasize the need for such guidelines.

My point of departure is this: observational work is unavoidable, but extracting conclusions from literatures dominated by it is difficult. Observational results are often highly sensitive to hard-to-detect errors, and so observational research frequently produces a swamp of contradictory findings. Only by producing quality guidelines that separate the wheat from the chaff can we expect observational work to play a meaningful role in policy analysis.

Such guidelines, however, will be substantially harder to develop than those for RCTs. Unlike RCTs, for example, observational researchers often need to employ a different method for each threat to quality, and they may even have to decide which of several options is optimal. They must also confront more explicitly the fact that an approach which improves one margin of quality may undermine another. Observational guidelines will therefore be significantly more complex, and analysts will have to wrestle with normative questions (such as how to trade off quality decisions) not present in RCT guidelines.

I start by explaining briefly why observational work is inescapable (and essential), and why it is thus necessary to design guidelines for it. I then turn to the difficulties that such guidelines face. Almost no work has been done on this issue, so my focus is not—cannot be—on providing answers, but rather on sketching out the basic concerns that must be addressed and some possibly fruitful avenues to consider.

1 Observational Research: Unavoidable, But Risky

That no guidelines exist for observational work is no mistake. Groups such as the Campbell and Cochrane Collaborations explicitly and intentionally focus solely on experimental and quasi-experimental approaches. Others, such as the American Evaluation Association, may be moving away from an "RCT-only" focus, but slowly.

This RCT myopia is unfortunate. First, there are a host of important policy questions that can only be addressed through observational methods. RCTs may be ethically or pragmatically infeasible. It is unethical for toxicologists to randomly poison people, and it may be political suicide for a mayor to randomize police deployment. Also, RCTs return only the mean difference in responses, and sometimes other measures, such as the variation, are of equal

[*] Associate Professor of Law, Fordham Law School.

[1] "Observational research" can take on many meanings. I use it here to refer to the quantitative, non-experimental work that makes up the bulk of empirical social science research, such as regression analysis.

or more interest (see, e.g., Heckman and Smith 1995). RCTs also have a hard time measuring heterogeneous effects. Plus, social scientists face the problem that humans adjust their behavior when they know they are being experimented on or monitored. And finally, RCTs—which take years to conduct—may be impractical due to time constraints.

And second, while unavoidable, observational work is riskier to use than experimental. If nothing else, it is remarkably more complex. This is not inherently bad, since technically sophisticated methods often lead to more accurate results. But the more complicated the approach, the more room for error—and observational results can be particularly sensitive to small errors. In other words, observational work is both intricate and delicate, a dangerous combination.

Essential and fragile: these reasons alone would justify developing observational guidelines. But there is a third issue that interacts toxically with these two. Observational work has never been easier or more popular. The increase in computing power, the skyrocketing growth in storage capacity, and the development of user-friendly statistics packages over the past thirty years all allow anyone to produce an observational study, regardless of formal statistical training. The result is an explosion in observational empirical analysis, but much of it of low quality—to the point that overall average quality is likely declining. Some of the low-quality work is the product of naïveté, but some of it is also the product of cynical manipulation by those who wish to poison the evidence base for personal, political, or financial ends.

At times, it feels as if a corollary to Newton's Third Law applies: for every observational empirical claim, there is an opposite (though not necessarily equal) claim. Without quality guidelines to sort through this quagmire, observational analysis will eventually be unable to provide meaningful guidance. To date, however, evidence based policy has effectively ignored observational work, and disciplines like the social sciences that rely on observational analysis have likewise ignored the evidence based revolution. I hope it is clear, in the short space provided, why both sides are in the wrong.

2 The Challenge of Designing Observational Guidelines

Designing observational guidelines raises numerous issues that are either absent from, or far less pressing on, RCT guidelines; much of the work on RCT guidelines thus provides little guidance. In this section, I want to consider three such issues. First, observational guidelines need to create clear definitions of quality, since the solution to one methodological problem may aggravate a different one. Second, these guidelines need to design methods to determine when various threats to quality must be addressed: since solutions have costs, they should not be imposed unless needed. And third, the guidelines must develop ways to tractably sort through the methodological complexity of observational work.

2.1 Defining Quality

At one level, defining quality is relatively straightforward. Unbiasedness (or internal consistency), representativeness (or external consistency), and efficiency (or precision) are widely accepted traits of high-quality work. There may be other goals—the American legal system, for example, often excludes highly probative evidence because of ethical violations—but these three margins are likely the main ones in scientific and policy settings.

The more challenging issue is how to balance these margins of quality. An approach that fixes one problem can aggravate another: instrumental variables, for example, reduce simultaneity bias, but do so by reducing precision as well. How can we compare the quality an unbiased but imprecise result to a biased but precise one? Economists, for example, often prioritize reducing bias, but it is not clear *a priori* that this is always the right path to follow.[2] It may not be feasible to establish clear ground rules in advance—it may even be impossible to clearly define "high quality" until mid-review, when the reviewer has a better sense the specific tradeoffs (for example, how much less precise certain types of less-biased estimates are)— but reviewers need to be conscious of these sorts of quality interactions when deciding what is "optimal."

This need to define quality tradeoffs, however, may actually—and surprisingly—point to an underappreciated strength of observational work. A key feature of the RCT is that it is a single device which targets almost all threats to quality. RCT guidelines are thus free to focus primarily on whether the procedures for an RCT are properly followed; whatever value tradeoffs are embedded in the RCT approach are generally not considered. That observational work cannot rely on a one-stop shop like the RCT frees it (at a cost, of course) to weigh various quality factors differently in different contexts.

2.2 Detecting the Presence of Threats

At least with respect to unbiasedness, the RCT targets almost all threats to quality. Randomization eliminates or mitigates omitted variable bias, functional form dependence, and self-selection (except perhaps for decisions made after treatment starts). Clinical assignment likewise eliminates or mitigates simultaneity bias, and by recording the data themselves clinicians hopefully minimize truncation or censoring and avoid the errors-in-data biases that can plague observational data gathered by others (such as government agencies). An experimenter thus does not have to ask if, say, endogeneity is a risk: if it is, the RCT's procedures protect against it, and if not he would have nonetheless used the RCT.

Observational researchers do not have this luxury. Many of the corrections they use come with costs or cannot be used in conjunction with other methodological approaches.

[2] For example, if the true effect size is 6%, which result is more useful: 4% ± 0.5% (biased, but precise) or 6% ± 8% (unbiased, but so imprecise that the estimate is statistically insignificant from zero)?

Interaction terms, for example, can reduce omitted variable bias, but they can complicate some responses to simultaneity bias. Analysts must determine what threats are present and choose their methods accordingly.

The need to define what threats are present clearly extends to quality guidelines as well. A study that unnecessarily controls for a harm that is not present is actually of lower quality than one that does not so control. Thus observational guidelines must develop empirically-validated criteria to correctly determine whether a particular threat is present, an extra step that RCT guidelines simply do not have to take.

In some cases, there are well-established tests to determine whether a threat exists; analysts can use the Goldfeld-Quandt or Breusch-Pagan tests to detect heteroskedasticity. But other threats may not lend themselves so easily to quantitative testing. Consider endogeneity. While there exist quantitative tests, such as the Granger Causality test, there are strong reasons to view them warily.[3] If no rigorous test exists—or if there is no consensus concerning what test to use—how can we determine whether a threat is present? Two options present themselves. The first is theory or intuition. The second is a literature review, as long as there is variation in the literature. If some studies control for a threat and others do not, variation (or its absence) in results may be informative. (If all studies use the same methodology, however, this method fails.)

Both these approaches are imperfect. Evidence based policy is motivated in large part by skepticism about theory and intuition, and using the literature itself to determine whether a threat is present undermines the goal of designing guidelines prior to the analysis. But these concerns do not imply that observational work should be ignored or that designing guidelines for it is futile—though they may reflect meaningful epistemic limitations to the evidence-based conclusions one can draw from observational work. Guidelines still have the ability to inject much-needed transparency and rigor into how knowledge claims are extracted from observational research. The claims may be more tentative, but they will be still be informative (perhaps in part *due to* their tentativeness).

2.3 Handling Complexity

Observational guidelines are layered like onions. Just to start, they must consider at least three margins of quality, just one of which—unbiasedness—faces about seven different threats. Here, I use as an example how to address just one of the threats to unbiasedness, namely endogeneity.

Researchers have developed numerous methods to treat endogeneity: quasi-experiments, regression discontinuity, instrumental variables, systems of equations, and other (lesser) approaches such as state-year trends. And so a guideline could simply ask (1)

[3] Though a detailed explanation is beyond the scope of this note, the basic intuition is that an endogenous relationship which bedevils analysis can similarly bedevil the test for its presence.

is endogeneity present? and (2) if so, does the paper use one of these approaches? But this seems insufficient. What is of interest is whether the paper uses the correct approach and uses it well—and this reveals another layer of criteria. For instrumental variables alone there are at least three sub-issues: Is the instrument exogenous? consistent? strong?[4]

Yet we can peel back even more layers. Guidelines could simply ask "is the instrument exogenous?" or they could specify more precisely how to establish exogeneity. There are at least three classes of tests to determine exogeneity (such as the over-identification test), and there are multiple options per class (such as the Sargan-Hansen or Bassman over-identification tests). Similarly, there are several tests to see whether the instrument in strong and numerous technical fixes to employ if not. The technical specifics are not important for the discussion here; the wide range of options is.

It is thus easy to see how quickly complexity can mushroom (to mix food metaphors). There are three prongs to quality, including unbiasedness. There are approximately seven threats to unbiasedness, including endogeneity. There are about five fixes for endogeneity, including instrumental variables. There are at least three facets to a good instrumental variable, including exogeneity. There are three or so tests for exogeneity, including the over-identification test, and there are at least three types of these tests. Other parts of the quality "tree" (to introduce more vegetation) exhibit similar branching.

Thus, observational guidelines need to tame this complexity if they are to be effective. This is no easy task, and it requires short-, medium-, and long-run goals. In the short run, it is imperative simply to centralize all the methodological options. Econometric textbooks may lay out the basics, but new approaches are developed regularly, and they are widely scattered. General solutions may be located in topical articles read only by a subset of a discipline (i.e., a new approach for handling self-selection may appear in a labor economics article) or in theoretical pieces not read by many practitioners.

And it is not enough for each field to undertake this alone: interdisciplinary collaboration is essential. Advances in political science may be helpful to economics or even epidemiology. Overcoming disciplinary boundaries, however, is not easy, and a centralizing association, such as the Campbell Collaboration, may need to assist the process. Observational methodology has been developing rapidly, and there appears to be little concerted effort to consolidate the advances (save some isolated examples, such as Imbens and Wooldridge 2009).

In the medium run, attention should shift to empirical validation. It is not just enough to assemble a list of methods: it is critical to understand when each is appropriate and what turns on the various options. Validation is growing in the medical sciences, but it is almost unheard of in the social sciences (Higgens and Green 2008). The methodological (over-) abundance observational researchers face implies that this will be a time-intensive process,

[4] The costs of instrumentation to representativeness and efficiency could be included here as well, though these relate to quality concerns other than unbiasedness.

but an essential one. Without it, researchers will often lack rigorous evidence about what methods are best for a particular situation. Validation will require within-literature assessments ("how do results produced using method *x* differ from those using *y* for question *z*?") followed by cross-literature analyses ("for what types of questions is method *x* superior to method *y*?").

In the long run, the goal is a master checklist. The number of methodological options is just too vast for any one researcher to keep them all in mind at all times. But while *some* prescription is needed, there also needs to be room for flexibility. It is essential to evaluate new approaches as they are released, and new work using old methods provides additional evidence about the effect of older techniques.[5] This points to yet another margin of complexity faced by observational, but not RCT, guidelines. The Cochrane and Campbell Collaborations call on analysts to update RCT reviews every few years to account for new substantive findings. Reviews of observational work will need to do this as well as update the *guidelines themselves* (and thus their assessments of all earlier studies) in light of methodological advances. The greater stability of RCT methodology spares RCT guidelines from this latter type of updating.

These are just the most central of concerns; numerous others exist as well, the discussion of which is beyond the scope of this paper. For example, guidelines are often rightly criticized for reducing quality to a single numeric score (Greenland 1994), but given all the margins of quality involved here a multidimensional approach may simply be impractical. And there is also the question of how to prevent an explosion in the number of competing guidelines, something that already afflicts the methodologically-more-simple RCT guidelines (see, for example, Jüni et al. 1999).

The task may appear daunting. But this should not discourage the development of observational guidelines—in fact, it should *motivate* us to develop them. Petitti (1994) puts it well. Though speaking about epidemiology, her statement rings true for all empirical disciplines:

> If epidemiologists cannot define what constitutes quality in non-experimental studies, how is it possible to do studies that we all agree have merit? If meta-analysis fails because quality is elusive, then all of non-experimental epidemiology fails for the same reason.

Exactly so. Regardless of how hard they are to design, quality guidelines for observational work are essential. And with more and more contradictory studies piling up, it is no longer possible to put off tackling the profoundly difficult issues they raise.

[5] Thus an additional concern with prescriptiveness. As noted above, evidence about methodological choices may ultimately come from the literature itself, but only if it is methodologically varied. Overly-prescriptive guidelines could stifle this essential variation.

3 Sources

Greenland, Sander. 1994. “Quality Scores are Useless and Potentially Misleading.” *American Journal of Epidemiology* 140: 300–301.

Heckman, James, and Jeffrey Smith. 1995. “Assessing the Case for Social Experiments.” *Journal of Economic Perspectives* 9: 85–110.

Higgens, JPT, and S Green (editors). 2008. *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.0* (updated February 2008). Available on-line at http://www.cochrane-handbook.org.

Imbens, Guido, and Jeffrey Wooldridge. 2009. “Recent Developments in the Econometrics of Program Design.” *Journal of Economic Literature* 47: 5–86.

Jüni, Peter, et al. 1999. “The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis.” *Journal of the American Medical Association* 282: 1054–1060.

Petitti, Diana. 1994. “Of Babies and Bathwater.” *American Journal of Epidemiology* 140: 779–782.