The Central Role of Rosenbaum and Rubin’s Seminal Work

Andrew J. Spieker  
Department of Biostatistics  
Vanderbilt University Medical Center  
Nashville, TN 37203, USA

Abstract

Rosenbaum and Rubin’s seminal work on the propensity score set the stage for decades of subsequent developments in causal inference methodology for use in observational studies. In this commentary, I discuss two specific aspects of their work with particular emphasis on how they have shaped my understanding of causal inference: (1) the propensity score as a data reduction technique, and (2) the importance of drawing parallels between the observational study and the randomized experiment.

Keywords: Balance, Causal Inference, Observational Studies, Propensity Score

1. Introduction

It is a pleasure to have the opportunity to discuss Rosenbaum and Rubin’s seminal paper, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” which set the stage for decades of progress in methods for observational data research (1983). In this brief commentary, I primarily seek to summarize how the key concepts of this paper have broadly shaped my way of thinking about (and teaching) causal inference, with emphasis on connections to other aspects of causal thinking that have emerged in the past few decades. Throughout, I will generally preserve the notation as it appears in the original manuscript, letting $X \in \mathbb{R}^K$ denote the pre-treatment covariate profile, $Z \in \{0, 1\}$ the exposure of interest, and $R \in \mathbb{R}$ the outcome of interest.

2. The key concepts

There are many ways to describe and explain the challenges associated with drawing causal conclusions from observational data. I have found that the most accessible introduction to those unfamiliar with the causal framework involves underscoring the differences in pre-treatment covariates between the treatment groups (that is to say that $p(X|Z = z)$ depends upon the value of $z$). Study of the propensity score instead considers variation in $p(Z|X = x)$ across subgroups $X = x$. Though by definition these are equally valid ways of conceptualizing the dependence of $X$ and $Z$, there are key advantages to anchoring theoretical exploration to the latter formulation as Rosenbaum and Rubin have done:

1. From a causal perspective, thinking in terms of $p(Z|X = x)$ is more “proper” in that the descendant is conditioned on the ancestor rather than the other way around.

©2023 Andrew J. Spieker.
2. Exploration of \( p(Z|X = x) \) gave rise to what is arguably the most widely used data reduction technique of modern causal inference (i.e., \( x \mapsto e(x) = P(Z = 1|X = x) \)).

3. If we believe ignorability holds given \( X \), consideration of \( p(Z|X = x) \) more easily sets the stage for thinking about the parallels between observational studies and conditionally randomized experiments.

I will further expound upon the second point in §2.1, and on the third point in §2.2.

2.1 The propensity score as a data reduction technique

The first four pages, roughly, of Rosenbaum and Rubin’s manuscript center on characterizing the propensity score as the fullest possible distillation of a covariate profile that can be used to achieve balance (in the sense of Rosenbaum and Rubin’s §1.2). By the bottom of p. 44, we have learned that a function of \( X \) produces balance if and only if that function either produces the propensity score itself or if some function can be composed with it to produce the propensity score—that is, the full set of such functions can be expressed as:

\[
\mathcal{B} = \{ b : \mathbb{R}^K \rightarrow \mathbb{R}^M \text{ such that } \exists f : \mathbb{R}^M \rightarrow \mathbb{R} \text{ with } e = f \circ b \},
\]

with the non-trivial (i.e., dimension-reducing) balancing scores being represented by choices of \( b \) with \( 1 \leq M < K \). The result that follows (Theorem 3) is essentially to say that if \( X \) represents a set of pre-treatment covariates such that treatment is as if randomly assigned within subgroups defined by \( X \), so too will treatment be as if randomly assigned within subgroups defined by \( b(X) \), for any choice of \( b \in \mathcal{B} \). Since this conclusion is valid for any balancing score, the salient information about the (possibly high-dimensional) \( x \) is, at least in theory, carried fully by the scalar-valued \( e(x) \). Rosenbaum and Rubin go on to develop the asymptotic basis for matching, stratification, and adjustment (based on any balancing score), and then provide justification for these approaches for implementation in finite samples.

As our technological resources have grown over the years, so too has our ability to develop progressively more computationally intensive methodology. These advancements have influenced the methods and mechanisms of data reduction associated with estimation of the propensity score, and have broadened their applicability. We have seen, for instance, research focused on ways to better estimate individualized propensity scores using machine learning approaches, with applications appearing in a variety of fields including social and health sciences (Ferri-Garcia and del Mar Rueda, 2020; Goller et al., 2020; Linden and Yarnold, 2016; Lee et al., 2010; Li et al., 2016). Further, we have seen generalizations and extensions of the propensity score to accommodate ordinally and continuously assigned exposures (Joffe and Rosenbaum, 1999; Lu et al., 2001; Imai and Van Dyk, 2004; Hirano and Imbens, 2004).

One the other hand, we have seen research emerge on some of the potential pitfalls associated with the propensity score as a dimension-reducing technique. In recent years, the issue of whether the propensity score distance adequately characterizes similarity between treated and control units for the purposes of determining matches has been vigorously debated. From the standpoint of large-sample validity under the usual causal assumptions, the theory is on the side of Rosenbaum and Rubin. However, small-sample behavior is
also important to explore, particularly when there are regions of poor covariate overlap and multiple matching procedures from which to choose. King and Nielsen (2019) severely criticize the use of the propensity score for this purpose (though, notably, not for other purposes such as weighting or adjustment), heuristically explaining that a fundamental weakness of propensity score matching lies in its attempt to approximate a fully randomized experiment. They argue, instead, that approaches to match on scale-invariant covariate distance should be leveraged to mimic a fully blocked randomization procedure. On the other hand, Ripollone et al. (2018) seem to argue that the circumstances in which propensity score matching is truly problematic in practice is sufficiently narrow (for instance, when extreme calipers are used) so as not to abandon the idea altogether. Guo et al. (2018), too, take a more moderate view on propensity score matching, arguing that no one approach will dominate in all settings. Rosenbaum and Rubin acknowledge that it is not necessarily always desirable to maximally distill the information provided by $X$ (the example provided is when one seeks to estimate covariate-specific treatment effects).

2.2 From observational data to randomized experiment

Although Rosenbaum and Rubin specifically set out to deal with the case of observational data, they facilitate their presentation of the propensity score by referring back to the randomized experiment throughout their manuscript. The randomized experiment is, of course, the gold standard by which causal associations can be ascertained due to procedurally-ensured absence of systematic confounding. One of the primary ideas formalized in their work is that when an observational study can be realized as a conditionally randomized experiment, it can be analyzed as such and causal claims will be warranted.

In my experience, it is often conceptually helpful—both when applying causal methods to real-world data and when teaching the theory and methods to students—to draw parallels between an observational study and a randomized experiment as Rosenbaum and Rubin have done. A fully and unconditionally randomized trial disrupts the flow of information from $X$ to $Z$ that might otherwise exist in an uncontrolled population, axiomatically and by study design (i.e., without untestable modeling assumptions). Commonly implemented methods involving the propensity score (e.g., adjustment by conditioning, matching, and weighting) all seek to disrupt the flow of information from $X$ to $Z$, although in ways distinct from one another and from the randomized trial (see Figure 1). Below is a heuristic description of how each comparator approach breaks the flow of information along the pathway $X \rightarrow Z$, with a very brief summary of some special considerations that can guide informed methodological choices.

1. Covariate adjustment for $e(X)$. Positing a proper form for $E[R|Z = z, e(X) = e_0]$ disrupts the flow of information from $X$ to $Z$ by ensuring that the target parameter reflects a comparison of subgroups with covariate profiles that carry the same salient information about treatment. As noted by Rosenbaum and Rubin, the model must adequately capture the propensity-outcome relationship to consistently estimate the target parameter (of note, Zhou et al. (2019) recently proposed a principled approach incorporating penalized splines on propensity scores for estimation of causal effects). Importantly, covariate adjustment alone will not target a population-level association if the measure of contrast is not collapsible.
2. Matching on $e(X)$. If the matching algorithm is successful, this direct approach results in a sample in which the covariates are (at least approximately) balanced between treatment groups, thereby breaking the dependence of $X$ and $Z$ in the final sample. I concur with Rosenbaum and Rubin’s comments that this approach tends to be readily accessible to scientific collaborators (p. 48). Compared to the propensity adjustment approach, near-violations to the positivity assumption make themselves more apparent and can be controlled with calipers (at the cost of inducing bias). Indeed, if all treated units are retained and the matched controls are drawn from a large pool of candidates, the target parameter must be realized as the treatment effect “on the treated” if there is treatment effect heterogeneity. Further note the considerations on matching discussed in §2.1.

3. Weighting involving $e(X)$. Although not specifically addressed by Rosenbaum and Rubin’s 1983 paper, I would be remiss not to include inverse probability-of-treatment weighting (IPTW) as one of the most well-known uses of the propensity score to estimate causal effects (Robins et al., 2000). This approach seeks to generate sample balance (in the weighted sample), but differs from the matching approach in that it targets a causal effect in the full population. Like matching, near-violations to the positivity assumption make themselves more apparent in this approach as compared to propensity adjustment and can be controlled by approaches such as weight truncation (also at the cost of introducing bias). Further, this use of the propensity score has advantages when dealing with non-collapsible measures of association, and also generalizes readily to causal inference involving longitudinal treatments.

Indeed, methods for estimating causal effects not based on the propensity score can also possess parallels to the randomized experiment. Instrumental variable methods (Angrist et al., 1996; Baiocchi et al., 2014) have grown in popularity for use in observational data as certain classes of causal parameters can be identified even when the ignorability assumption is not thought to hold (e.g., when variables contained in $X$ go unmeasured), provided there exists a cause of $Z$ (call it $W$) that is conditionally independent of $R$ given $Z$ (McClellan et al., 1994; Brookhart and Schneeweiss, 2008; Lorch et al., 2012). In this sense, $W$ serves as a natural experiment and functions as a random treatment allocation process. Less obvious, in my view, is that instrumental variable methods can also be viewed as an alternative way to disrupt the flow of information from (possibly unmeasured) $X$ to $Z$. This is easiest to argue in the case where $W \in \{0,1\}$. Letting $\mathcal{P}$ denote a full population of interest partitioned into the four usual principal strata, the local average treatment effect can be expressed as:

$$
\Delta = \mathbb{E}[R_{w=1} - R_{w=0} | Z_{w=1} - Z_{w=0} = 1],
$$

of course invoking the monotonicity assumption of Imbens and Angrist (1994). This target parameter can be realized as the average causal effect among a suitable subset of the population $P_0 \subset \mathcal{P}$. The special property possessed by this sub-population is complete concordance between $W$ and the potential outcomes $Z^W$. Since the exclusion restriction precludes the flow of information between $X$ and $W$, there can logically be no flow of information from $X$ to $Z$ among those in $P_0$, the restricted population.
Commentary on Rosenbaum and Rubin (1983)

Figure 1: Causal directed acyclic graphs representing the relationship between study variables. Panel (A) reflects the ideal situation represented by the (unconditionally) randomized experiment, in which pre-treatment causes of the outcome have no systematic association with the exposure. Panel (B) can be used to represent both the conditionally randomized experiment and the observational study in which ignorability of treatment is achieved given $X$. Many causal inference approaches specifically seek to either directly or indirectly disrupt the flow of information along the path $X \rightarrow Z$ (Panel (C)) in order to mimic the randomized experiment. One of Rosenbaum and Rubin’s essential contributions is the finding that the (possibly high-dimensional) covariate profile $X$ can be distilled down to the more coarsely defined propensity score, $e(X)$ (conditional on which the flow of information from $X$ to $Z$ is sufficiently disrupted to identify causal effects).

In summary, one key property shared by the most widely used causal inference methods suitable for single-time treatments is the disruption of the systematic flow of information along the pathway from $X$ to $Z$. There is diversity in the mechanisms available to accomplish this goal—and, in turn, identify causal effects of $Z$ on $R$—which include study design, defining a contrast of interest conditionally by modeling, exclusion of some controls in a sufficiently large reservoir, re-weighting to achieve balance, and shrinking the target population to achieve synergy between the instrumentally-randomized exposure and the exposure actually observed.

3. Additional comments and concluding thoughts

Speaking more broadly for a moment than the specific work presented in Rosenbaum and Rubin’s 1983 manuscript, the potential outcomes framework for causal inference as a whole has been highly influential in my own methodological research. I believe that the reason this framework has endured is that one must carefully articulate assumptions and define target parameters at the outset. In this sense, the particulars of the estimation procedures are
very much secondary to the assumptions. Further, as we see from Rosenbaum and Rubin: from one set of assumptions, many approaches emerge.

The field will continue to produce developments, expansions, and refinements to the existing set of causal approaches in the decades to come, and we should look to Rosenbaum and Rubin (1983) as an example, particularly for the strong emphasis on articulating assumptions and acknowledging the relative advantages and limitations of procedures and approaches. Any identifiability assumptions that cannot be assessed empirically or guaranteed by study design should be defended *a priori*. Where possible violations to untestable assumptions are suspected, sensitivity analyses are highly useful to assess robustness of scientific conclusions (Rosenbaum, 1987; Lin et al., 1998; Spieker et al., 2022). Empirical studies can also augment our understanding of a model’s robustness to untestable assumptions. However, where sensitivity approaches are undeveloped or infeasible, and where empirical studies demonstrate non-robustness, the next line of attack is generally to try to trade in the problematic assumption(s) for others that better hold water. In fact, it is for this reason that the first causal methods to which I was exposed were developed primarily in the econometrics literature (e.g., Heckman (1978); Freedman and Sekhon (2010)). Specifically, the challenge at hand was to develop and employ methods to estimate causal effects that rely on neither an assumption of ignorability nor the availability of an instrumental variable. Causal parameters generally cannot be nonparametrically identified in such circumstances; identification instead comes from correct specification of a likelihood function for the data (Maddala, 1983; Spieker et al., 2021). All this is to say that Rosenbaum and Rubin’s work is a key piece among several from that era that impresses upon me the importance of critical consideration of assumptions. To blindly apply methodology without doing so would be a gross misuse of their theoretical work.

At the time of this writing, most of my graduate students’ research areas (both applied and methodological) rely heavily on the potential outcomes framework for causal inference. Rosenbaum and Rubin’s 1983 paper is therefore routinely one of the very first I assign as a reading. What I don’t tell my students (although I suppose they will find out if they read this commentary) is that half the reason I assign it to them is that it provides me an excuse to go back and read the manuscript myself—something from which I derive benefit each time, and evidently without diminishing returns.
References


