Propensity Score Methods in Nursing Research
Take Advantage of Them but Proceed With Caution

Wei Pan ▼ Hailian Bai

Intervention research on health outcomes is important to advancing the nursing science. Using randomized controlled trials (RCTs) to estimate intervention (or treatment) effects is optimal for this purpose. Unfortunately, for practical or ethical reasons, RCTs are often not feasible, and thus, researchers often rely on observational or non-RCT data to estimate treatment effects. Such practice poses a threat to the validity of treatment effect estimation, because the lack of randomization introduces selection bias in observational and non-RCT data. Since the inception of Rosenbaum and Rubin’s (1983) seminal work on propensity score methods for reducing selection bias, numerous studies have been published in the social, behavioral, and medical literature in applying propensity score methods to observational studies and non-RCTs.

A propensity score is defined as the conditional probability of a subject (e.g., patient) being assigned to a treatment group given a set of observed covariates (or confounders) such as age, gender, race/ethnicity, and health status (Rosenbaum & Rubin, 1985). Propensity score methods are a set of statistical procedures for using propensity scores to balance the distributions of observed covariates between the treatment and control (or comparison) groups with the aim of reducing selection bias. Such covariate balance enables a direct comparison between the treatment and control groups, similar to RCTs; therefore, applying propensity score methods can increase the validity of treatment effect estimation when observational or non-RCT data are used (Bai, 2011b; Pan & Bai, 2015b; Rubin, 2008). Propensity score methods usually consist of four basic steps: (a) selecting covariates; (b) estimating propensity scores; (c) matching, weighting, or stratifying subjects on/using propensity scores; and (d) conducting the intended outcome analysis (Pan & Bai, 2015a).

Although propensity score methods are widely used in social, behavioral, and medical research, few publications related to propensity score methods are found in nursing research. For example, a search in Web of Science using “propensity score” as a topic keyword showed that, of the 14,561 articles related to propensity scores published since 1983 in Web of Science categorized journals, only 48 of them (0.33%) were published in nursing journals. The dearth of applications of propensity score methods to nursing research suggests that propensity score methods may not be well known to nursing researchers. Random assignment is often infeasible or unethical in healthcare settings. Thus, using observational or non-RCT data that include information about interventions or treatments, such as the electronic health records, provides nursing researchers the advantage of applying propensity score methods to their research for advancing the nursing science (Samuels, McGrath, Fetzer, Mittal, & Bourgoine, 2015).

In this issue of Nursing Research, it is encouraging to see a methodological study by Schroeder, Jia, and Smaldone (2016) that gives an overview of three propensity score methods (matching, stratification, and weighting) and compares their covariate balance using observational data extracted from the electronic health records that contain information about a school nurse-led obesity program. Results from this study reveal that the performance of the three propensity score methods in covariate balance varied, with propensity score matching providing the best covariate balance. The authors recommend that various propensity score methods be examined for covariate balance before conducting the intended outcome analysis.

As the first methodological study in the nursing literature to examine the effectiveness of various propensity score methods for improving covariate balance and the fourth such article in the history of Nursing Research (the other three are Coffman & Kugler, 2012; Eckardt, 2012; Qin, Titler, Shever, & Kim, 2008), Schroeder et al.’s (2016) study provides additional helpful information to nursing researchers about the application of propensity score methods using observational or non-RCT data to increase the validity of treatment effect estimation. Their study extends the literature on comparison with propensity score matching techniques only (Austin, 2014; Bai, 2011a; Gu & Rosenbaum, 1993) to all three major types of propensity score methods. In addition, the finding from their study that propensity score matching outperforms propensity score stratification is consistent with those from Austin (2007), Austin and Mamdani (2006), and Austin and Schuster (2016). Yet, this study also found that propensity score weighing does not perform better than propensity score matching or stratification, which is inconsistent with other comparative studies (Austin, 2010; Austin & Schuster, 2016; Kurth et al., 2006; Lunceford & Davidian, 2004; Stone & Tang, 2013; Weeks et al., 2015). The inconsistency could be due to the differences...
in data and criteria used to evaluate the performance of propensity score methods. Lastly, Schroeder et al. (2016) discussed two important caveats for researchers to consider when applying propensity score methods: (a) the selection of propensity score methods should be guided by statistical theory and the nature of the research design, and (b) multiple propensity score methods should be tested for ability to improve covariate balance before proceeding with outcome analysis for estimating treatment effects.

It is also worth noting that for evaluating the performance of propensity score methods in achieving covariate balance, Schroeder et al. (2016) simply counted the number of covariates with significant differences between the treatment and control groups removed solely based on individual statistical hypothesis tests. This procedure is not recommended in the literature (Imai, King, & Stuart, 2008) because (a) hypothesis tests are affected by factors, such as sample size and variance, other than covariate balance; (b) covariate balance is a characteristic of the sample, not some hypothetical population to which hypothesis tests infer; (c) the evaluation of the performance of propensity score methods on covariate balance should be based on global balance for all of the covariates as a whole not each individual covariate; and (d) some covariates with small initial differences between the treatment and control groups likely sacrifice their own balance to achieve global balance and, therefore, may produce “quite unstable” balance for themselves (Rosenbaum & Rubin, 1985, p. 36). Alternatively, many other commonly used, more informative statistics and graphics for assessing the performance of propensity score methods on covariate balance are available, such as standardized differences, percent reduction in bias, Q-Q plot, and the Love plot (Ahmed et al., 2006; Cochran & Rubin, 1973; Pan & Bai, 2015a; Pattanayak, 2015; Rosenbaum & Rubin, 1985). Another point to note is about the criterion used for assessing the performance in covariate balance specifically for propensity score stratification: “if a statistically significant difference existed in at least one strata, the confounder was counted as a significant difference even if it did not differ within the other four strata” (Schroeder et al., 2016, p. 469). Such a criterion is too conservative; therefore, the weighted average across all five strata is commonly used to evaluate covariate balance (Pattanayak, 2015; Rosenbaum & Rubin, 1984).

Provided that propensity score methods are applied correctly, they can improve the validity of treatment effect estimation using observational or non-RCT data. Therefore, apart from the two fundamental assumptions of propensity score methods (the strong ignorability in treatment assignment and the stable unit treatment value assumption) discussed in the literature (Pan & Bai, 2015a; Rosenbaum & Rubin, 1983; Rubin, 1980, 1986) and the two caveats discussed in Schroeder et al. (2016), it is beneficial for researchers to be aware of some other cautions in the use of propensity score methods. Those additional cautions are that propensity score methods:

- can reduce overt selection bias from observed covariates but cannot reduce hidden bias from unobserved covariates;
- need a sufficient overlap (or common support) of the distributions of covariates between the treatment and control groups; and
- do not have a built-in mechanism for handling unwanted covariates that are related only to the treatment assignment but not the outcomes, which may increase the error variance of the estimated treatment effect (Brookhart et al., 2006).

To correctly use propensity score methods, researchers need to proceed with caution. First, to reduce overt selection bias from observed covariates as much as we could, all observed covariates related to both the treatment assignment and outcomes, based on theory and prior research, should be considered in propensity score estimation models. It is essential for researchers to present comprehensive information on covariate selection to justify the inclusion of covariates in propensity score estimation models. It is also necessary to conduct sensitivity analysis to test the model sensitivity to hidden bias from potential unobserved covariates (Groenwold & Klungel, 2015; Li, Shen, & Li, 2015). Second, to improve the efficiency of creating a matched control group, a large sample should be used for increasing common support between the treatment and control groups (Rubin, 1997). Finally, a set of covariates, thoughtfully selected by assessing their relationships with the treatment assignment and the outcomes, will help attenuate the influence of unwanted covariates on the estimation of treatment effects.

In short, propensity score methods are an effective statistical tool for reducing selection bias in observational and non-RCT data. Because of the practical or ethical barriers to conducting RCTs in healthcare settings, applying propensity score methods to observational and non-RCT data, such as the electronic health records, is an advantageous alternative to using RCT data to estimate treatment effects in nursing research. To obtain valid treatment effect estimates from observational and non-RCT data, researchers need to proceed with caution and apply propensity score methods appropriately.

Wei Pan, PhD, is Associate Professor and Director of Research Design and Statistics Core, School of Nursing, Duke University, Durham, North Carolina.

Haiyan Bai, PhD, is Associate Professor, Department of Educational and Human Sciences, University of Central Florida, Orlando.

The authors have no conflicts of interest to report.

Corresponding author: Wei Pan, PhD, School of Nursing, Duke University, 307 Trent Dr., DUMC 3322, Durham, NC 27710 (e-mail: wei.pan@duke.edu).

Accepted for publication August 31, 2016.

DOI: 10.1097/NNR.0000000000000189

RESPONSE

We thank Drs. Pan and Bai for their thoughtful contribution to our discussion about propensity score methods. As noted in the commentary and discussed in our manuscript, propensity
scores may be useful in nursing research, yet they must be applied with attention to their inherent limitations. We agree with Drs. Pan and Bai that our findings regarding propensity score weighting were unexpected. In our manuscript we posit that this unexpected finding may be due to the fact that receipt of our treatment of interest was rare (5% enrollment rate). In addition, as we discuss in the manuscript and as noted by Drs. Pan and Bai, we used a conservative method to measure improvement in confounder balance for propensity score stratification; other researchers may want to use a less conservative method. Lastly, as we acknowledge in our limitations, we measured confounder imbalance using a straightforward count of significant confounder differences; in future studies, nurse scientists may want to consider additional methods for assessing propensity scores’ impact on covariate balance.

We hope that our manuscript and Drs. Pan and Bai’s commentary will contribute to increasing discussion about propensity score methods in nursing science. Given the relative dearth of publications on propensity scores within the nursing field, there is opportunity for nurses to engage in a more robust conversation about these methods. With cautious use and attention to their inherent limitations, propensity score methods may provide a useful tool for nurse scientists in evaluation of non-randomized treatments.

Krista Schroeder
Haomiao Jia
Arlene Smaldone

REFERENCES


