The Uses of Propensity Scores in Randomized Controlled Trials

Travis Loux
Department of Epidemiology and Biostatistics
Saint Louis University College for Public Health and Social Justice
St. Louis, MO 63104, USA

Yi Huang
Department of Epidemiology and Biostatistics
Saint Louis University College for Public Health and Social Justice
St. Louis, MO 63104, USA

Abstract
Propensity scores are a dimension-reduction technique used to quantify the differences between treatment groups. Though propensity scores were developed to address the issue of confounding in observational studies, they have also proven useful in randomized controlled trials where confounding is structurally absent. When applied to randomized controlled trials, propensity scores can ensure balance between groups at the time of randomization, account for chance imbalances in observed randomization, and generalize target results to target populations. In this article, we review propensity score methodology developed for randomized trials with these goals.

Keywords: Covariate Balance, Generalization, Randomization, Statistical Efficiency, Transportability

1. Introduction
The propensity score was introduced by Rosenbaum and Rubin as a tool for addressing confounding in observational studies (Rosenbaum and Rubin, 1983). Since the introduction of the propensity score, it has been suggested for use as a sub-classification, matching, and weighting tool in attempts to balance the distribution of covariates between treatment groups. The propensity score has been the motivating use case behind a particularly deep literature around matching methods (Austin, 2008; Stuart, 2010). While the vast majority of methodological and applied research on propensity scores has focused on observational studies, there is a growing literature considering their use in randomized controlled trials (RCTs).

The theory behind RCTs rests on the structural independence of treatment assignment and pre-treatment covariates, as well as the randomization distribution resulting from hypothetical randomizations of treatment assignment (David, 2008). Under a hypothesis of no intervention effect, the observed individual outcome under one treatment assignment is equal to the unobserved outcome under the alternative assignment. Thus, an analyst can “re-randomize” treatment assignment and compute a contrast between the hypothetical treatment and control group. Repeated for every possible treatment allocation, or a large
representative sample of them, the simulated contrasts form a distribution of potential contrasts under the hypothesis of no effect against which an observed outcome can be compared and a p-value computed.

The use of covariate adjustment in the analysis of RTCs has been a hotly debated issue for decades (Pocock, Assmann, Enos, and Kasten, 2002; Senn, 1989). Because treatment assignment is independent of covariates, confounding will usually be minimal and accounted for through poorly balanced assignments in the randomization distribution. In this case, a crude contrast such as the difference in means or the difference or ratio of proportions is an unbiased estimate of the causal effect of treatment. However, covariate adjustment can produce more accurate and precise causal estimates (Pocock et al., 2002; Senn, 1989). A valid practical concern arises in that once one allows for covariate adjustment, the various choices of variable selection and model specification allow for far more flexibility and data dredging than a crude comparison might, opening up analyses to risks of multiple testing, over-fitting, and p-hacking (Pocock et al., 2002). Pre-registration of analytical methods has been a popular suggestion for taking advantage of the benefits of covariate adjustment while trying to mitigate the risks opening such doors poses (Nosek, Spies, and Motyl, 2012).

A review of the literature will show three main applications of propensity scores in RCTs: treatment allocation, outcome analysis, and generalization of results. In all these areas there have been conversations in the RCT literature dating back decades, though the inclusion of propensity scores is a relatively new addition.

2. Propensity Scores in Treatment Allocation

Though simple randomization of treatment allocation is enough to elicit causal inferences, there have been numerous suggestions for increasing its efficiency. A very early development was blocking, in which study subjects are stratified on some characteristic and randomization occurs within each stratum, ensuring balance on the blocked characteristic (David, 2008). In addition, various types of “biased coin” designs were developed to ensure global balance or balance within pre-defined strata (Efron, 1971). In a biased coin design, the probability of allocation to treatment changes from one subject to the next depending on previous allocations. For example, if the goal of a design is to randomize one-to-one treatment to control, but midway through the trial only 30% of subjects have been assigned to treatment, the biased coin approach increases the probability of assignment to treatment for the remaining subjects. Once balance has been reached, allocation probabilities can return to their original settings and continued to be monitored. Even with changing probabilities of treatment assignment, balanced coin designs can provide unbiased effect estimates if the assignment probabilities are taken into account (Baldi Antognini and Zagoraiou, 2011). One limitation of blocking and balanced coin designs is the relatively few number of covariates they can handle. This issue is analogous to the curse of dimensionality in matching within observational studies. Like in observational studies, the propensity score presents a potential solution.

Loux (2013) developed propensity-biased allocation (PBA), which allocates the \((n+1)^{th}\) subject to the treatment with probability of \(\pi_{n+1} = \pi(p_{n+1})\), where \(p_{n+1}\) is the estimated propensity score for the \((n + 1)^{th}\) subject from a logistic regression fit on the previous \(n\) subjects using observed baseline covariates. The function \(\pi(p)\) is assumed to be non-
Propensity Scores in Randomized Controlled Trials

Increasing function with $\pi(0.5) = 0.5$ and suggested to be symmetric about (0.5, 0.5). PBA performs well without discrete variables or modeling assumptions.

Alternatively, Loux (2015) proposed propensity-constrained randomization (PCR), which summarizes the balance between treatment groups by variance of the empirical propensity score, shown to be a function of the differences in group covariate means. PCR provides a set of well-balanced treatment permutations to form contrasts. The resulting randomization distribution has smaller variance than the randomization distribution obtained through pure randomization.

Incorporating propensity scores in the assignment of treatment can help improve the efficiency of inferences from small RCTs. By reducing the probability of an extremely unbalanced sample, these procedures ensure that balance is more closely achieved, at least among the observed variables. The result is standard errors which are more closely tied to variation due to randomness in assignment and less to accounting for potential imbalance in important covariates.

3. Propensity Scores in Outcome Analysis

Covariate adjustment can improve estimates from RCTs in two ways. First, it may help adjust for minor imbalances in the observed randomization, yielding more accurate effect estimates. Second, covariate adjustment can account for variation in the outcome, reducing the residual variance and standard errors of effect estimators even if no confounding is present, yielding more precise estimates. In addition to concerns about over-fitting and p-hacking, covariate-adjusted effect estimates target a different parameter - the conditional average treatment effect - than crude comparisons (Greenland, Robins, and Pearl, 1999). For non-collapsible contrasts, such as the odds ratio, the conditional average treatment effect may be a different numeric value than the average treatment effect. The difference between these two parameters may not be due to confounding or imbalance between treatment groups (Greenland, 1996). Ultimately, observed differences between the crude and covariate-adjusted effect estimates can cause hesitancy in interpreting the results of a trial and are compounded if the quantitative differences become qualitative in terms of statistical significance or direction of effect, as seen, for example, in Simpson’s Paradox (Greenland et al., 1999).

Regression or ANCOVA methods could provide us a solution to correct imbalance from randomization through reducing the standard error of the effect estimator, though they are sensitive to the model misspecification. Another limitation of ANCOVA is that one may search for a model specification that provides the maximal overall treatment effect and thus jeopardize the objectivity of statistical analysis (Pocock et al., 2002).

Inverse probability of treatment weighting (IPTW) is one way to use propensity scores to balance subjects’ characteristics in the treatment and control groups by weighting each individual with the inverse probability of receiving his/her actual treatment. Weights are calculated for each individual as $\frac{1}{P_S}$ for the treatment group and $\frac{1}{1-P_S}$ for the control group. IPTW was demonstrated asymptotically equivalent to the efficient ANCOVA estimator (Shen, Li, and Li 2014) while limiting concern of model misspecification; however, it should be used with caution when the sample size is limited ($N \leq 150$).
To improve the precision without compromising objectivity, Shen, Li, and Li (2014) propose a two-stage estimation procedure and suggest that the IPTW weights for covariate adjustment should be determined before collecting the outcome data to reduce the possibility of model selection bias. This method prohibits the analysts from seeing the outcome data and the baseline covariates in the same data set, reducing the possibility of selecting favorable models through examination of the relationship between the covariates and the outcome.

Zeng, Li, Wang, and Li (2021) proposed the overlap weighting (OW) approach, which changes the weights from the reciprocal of the estimated probability of being assigned to the observed arm (IPTW) to the probability of being assigned to the opposite arm (OW): \( \frac{1}{1 - \frac{1}{PS}} \) for the treatment group and \( \frac{1}{\frac{1}{PS}} \) for the control group. OW completely removes imbalance, attaining the same semiparametric variance lower bound as the ANCOVA/IPTW estimator and consistently outperforming IPTW in finite samples. Yang, Li, Thomas, and Li (2021) extended OW to subgroup analyses and compared OW to a full-interaction ANCOVA-S model on the overall sample with baseline covariates, treatment indicator and their interactions with the subgroup variables. They showed that ANCOVA-S estimator is as efficient as weighting methods with full interaction, when the ANCOVA model is correctly specified, while OW estimator with propensity score estimated from a full-interaction model on the treatment indicator outperforms ANCOVA-S under small subgroup sample size, and/or when the ANCOVA is misspecified.

In observational studies, propensity scores are used to analytically adjust for differences between treatment and control groups, often by matching or weighting. Though confounding is by design not a concern in randomized trials, imbalance will likely happen and has historically been adjusted for by stratified analyses or regression modeling. The use of propensity score-based analyses for these purposes allows for a semi-parametric analysis that can provide more fine-tuned adjustment than stratification but is less reliant on modeling assumptions than regression approaches. More recent work has incorporated propensity scores into assessing treatment effects within subgroups and effect modification, highly relevant questions in a time of personalized and precision medicine.

4. Propensity Scores in Generalization

Selection of subjects into RCTs is a well-known concern to the external validity of study results (Kern, Stuart, Hill, and Green, 2016). In the medical setting, access to institutions conducting trials (e.g., academic medical centers) as well as historical and social barriers to healthcare access generally, have resulted in study samples that tend to be more white, male, and middle-to-upper income than the general population, leading to an evidence base biased towards these populations. As personalized/precision medicine gains traction, some worry this imbalance in evidence may lead to even greater racial health disparities as the most recent advances may only be available to historically privileged groups over-represented in most medical RCTs (Borrell et al., 2021). Additionally, there may be logistical reasons why an RCT must take place in a relatively small geographic area before the intervention can be expanded to the general population. In such cases, the preliminary region is likely to be unrepresentative of the rest of the target site and results from initial investigations may not generalize well to a broader population (Kern et al., 2016).
To extend results of RCTs to the target population, (Cole and Stuart, 2010) propose using inverse probability-of-selection weights, which standardize observed trial results to a specific target population to generate inferences about the target population. Another strategy to obtain generalizability is to use the existing data to assess the comparability of RCT participants to the target population. Assuming the individual level data are available for population, or at least for a representative sample from that population, Stuart, Cole, Bradshaw, and Leaf (2011) proposed a “propensity-score-based” metric to quantify the similarity of the participants in a randomized clinical trial and a target population by providing summary measures of representativeness with respect to observed pre-treatment characteristics.

Rosenman, Owen, Baiocchi, and Banack (2022) proposed multiple methods to combine the information in a large biased observational database (ODB) with a small RCT from which treatment effect estimates would have high variance. These procedures use propensity scores developed in the observational database to create strata capturing heterogeneity in the causal effects. The “spiked-in” estimator proposed by Rosenman spikes the RCT data into the appropriate strata determined by the ODB, using the propensity score model from the ODB to estimate the propensity score for the RCT units had they come from the ODB. This estimator could improve the ODB estimator by increasing the number of treated units in the low-propensity edge bins and the number of control units in the high-propensity edge bins. An overall effect estimate is calculated as a weighted average of the spiked strata effects. Rosenman also considers a “dual-spiked” estimator by stratifying jointly on the propensity score and prognostic score (Hansen, 2008) to better account for treatment effect heterogeneity. The spiked-in estimators perform well when the covariate distributions in the two databases do not differ sharply; otherwise, the estimators could perform worse than using the ODB without the RCT. As an alternative, Rosenman proposes a “dynamic weighted average” estimator, which uses weights to combine the ODB-only estimate of the treatment effect in stratum k and the RCT-only estimator in stratum k. Simulations show this estimator is robust to differences in covariate distributions. At any rate, the inability of overcome drastically different covariate distributions in the RCT and ODB may not be limitation of propensity scores, but more an indication that such generalization is not appropriate given the lack of representation in the RCT database. This would be akin to poor propensity score matching or weighting diagnostics being a sign of poor balance and comparability rather than an indictment of the methodology itself.

In the previous sections, propensity scores were used as a tool to increase the efficiency of already unbiased effect estimates. In attempting to generalize results of a trial to a broader population, propensity scores are directly addressing an issue of bias. This is similar to their original intention in observational studies, though the source of the bias is different (selection instead of confounding). The propensity score, or propensity score-motivated approaches, provide a way to compare individuals in a trial to those in the general population, analogous to the comparison of treated individuals to controls. Used in this context, propensity scores can give better estimates for effects one might expect if an intervention was made more broadly available.
5. Conclusion

While propensity scores were originally developed to address confounding in observational studies of causal effects, recent literature has shown that they are also helpful in randomized studies as well. Though confounding is not a concern in randomized trials, any realized randomization will likely have some imbalance between intervention groups. Propensity scores can be used to minimize this imbalance at the randomization stage, analogous to blocking, or adjust for between-group differences in the analysis of outcomes, analogous to covariance adjustment. Both uses of propensity scores can increase the efficiency of randomization studies, especially in small samples or in investigating subgroup effects. Propensity scores, or propensity-based tools, can also be used to account for selection bias into randomized trials in hopes of generalizing or translating evidence from a randomized trial to a broader population. While not required for the heavy lifting of confounding adjustment, a thoughtful use of propensity scores in randomized trials can lead to improved estimates in terms of both statistical efficiency and applicability of results.

A review of the current literature suggests directions for future investigation into the methodology and application of propensity scores in randomized trials. One overarching question is the utility of machine learning in estimating propensity scores. Because the assignment mechanism in RCTs is known, logistic regression incorporating covariates used in determining treatment probabilities will estimate the true propensity score and theoretically provide consistent effect estimates; however, in RCTs the propensity score is used to identify observed imbalances in treatment assignment rather than estimate the true assignment mechanism. The potential for machine learning to better identify imbalances and improve propensity score estimates in RCTs is unknown. Many of the proposed methods, though, show the greatest benefit in relatively small samples where the smoothing of traditional statistical methods is helpful and machine learning techniques designed for large-scale data sets may not work well in application. Additionally, more work can be done in complex RCT settings. All the methods described above assume, at least implicitly, two study arms - treatment and control - and most are interested in an average treatment effect. Generalization of these techniques to multi-arm trials and investigating effect heterogeneity could reveal additional benefits to using propensity scores in RCTs. Similarly, the methods above ignore the complications of non-compliance and attrition. Investigating their limitations in these settings, and how they can be incorporated into methods addressing these issues, will help improve the application of propensity score-based methods in real-world RCTs. While propensity scores in treatment assignment and outcome analysis could be considered alternatives to traditional methods with similar goals, we see the use of propensity scores in generalization as an area of research with the most potential for high-impact methodological development.

RCTs are praised for their lack of confounding, and it may seem propensity scores have no place in this area of statistical work. Recent work proves otherwise. Though RCTs are not subject to confounding per se, incorporating propensity scores in the design and analysis stages of an RCT can identify and address observed imbalances in treatment assignment, ultimately improving the statistical efficiency of the estimates obtained in these studies. Additionally, propensity scores have been shown effective in addressing one of the common limitations of RCTs: lack of generalizability through selective sampling. A growing body
of literature shows the impact propensity scores can have in these important studies and there are plenty of avenues for further methodological investigation laying the foundation for consequential clinical and scientific research.

References


