Propensity Score Analysis: Recent Debate and Discussion

Shenyang Guo *Xi'an Jiaotong University & Washington University in St. Louis*
Mark Fraser *University of North Carolina at Chapel Hill*
Qi Chen *Xi'an Jiaotong University*

ABSTRACT Propensity score analysis is often used to address selection bias in program evaluation with observational data. However, a recent study suggested that propensity score matching may accomplish the opposite of its intended goal—increasing imbalance, inefficiency, model dependence, and bias. We assess common propensity score models and offer our responses to these criticisms. We used Monte Carlo methods to simulate two alternative settings of data creation—selection on observed variables versus selection on unobserved variables—and compared eight propensity score models on bias reduction and sample-size retention. Based on the simulations, no single propensity score method reduced bias across all scenarios. Optimal results depend on the fit between assumptions embedded in the analytic model and the process of data generation. Methodologic knowledge of model assumptions and substantive knowledge of causal mechanisms, including sources of selection bias, should inform the choice of analytic strategies involving propensity scores.

KEYWORDS: endogeneity, observational studies, propensity score analysis, propensity score matching, selection bias

doi: 10.1086/711393

Propensity score analysis (PSA) is a class of statistical methods developed for estimating treatment effects with nonexperimental data and, more generally, for estimating conditional causality with observational data. Specifically, PSA offers an approach to program evaluation when randomized trials are infeasible or unethical, or when researchers need to assess treatment or causal effects from survey data, census data, administrative data, medical records data, or other types of observations where a counterfactual must be constructed. (For additional background on PSA methods and applications, see Guo & Fraser, 2015.) In the social and health sciences, researchers often face a fundamental task of drawing conditioned casual inferences from quasi-experimental studies. Analytical challenges in making causal inferences can be addressed by a variety of statistical methods, including a range of new approaches emerging in the PSA field.

Journal of the Society for Social Work and Research, Volume 11, Number 3. 2334-2315/2020/1103-0006/$10.00.

The PSA approach has been used in a variety of disciplines and professions, including epidemiology (e.g., Normand et al., 2001), medicine (e.g., Gum et al., 2001), psychology (e.g., Jones et al., 2004), social work (e.g., Guo et al., 2006), and sociology (e.g., Morgan, 2001). The method was the eighth most popular approach in social work research identified from a review of articles published in five leading social work journals from January 1, 2012, to December 16, 2013: Of 167 articles using at least one multivariate statistical method, eight (4.8%) used PSA, and PSA was the principal method used to address the endogeneity problem in the reviewed articles (Guo, 2015). (See the following section for the definition and further discussion of the endogeneity problem.)

Debates regarding PSA and other bias-reduction statistical methods have been robust (e.g., Guo & Fraser, 2015). Bias-reduction methods are needed when researchers conduct observational studies but cannot employ randomized experiments or when there is evidence to show that such experiments are compromised or failed. A recent paper by Gary King and Richard Nielsen (2019) challenged the statistical conclusion validity of findings based on propensity score matching. The issues King and Nielsen raised are relevant to researchers from all professions and disciplines: What are the statistical problems pertaining to propensity score matching (PSM), what are the best alternatives given the range of available propensity score methods and the range of data generation situations that researchers encounter, and what kinds of knowledge do researchers need when choosing and applying propensity score models?

We address these questions. For a general discussion of various propensity score methods, readers are referred to Rosenbaum (2002), Imbens and Wooldridge (2009), and Guo and Fraser (2015). In this article, to provide a larger context, we begin with an overview of available PSA models, including matching on propensity scores. We then review the problems with matching suggested by King and Nielsen. To extend and examine findings of King and Nielsen's simulation studies, we present results of a Monte Carlo study using eight statistical models, including an ordinary least squares regression that does not correct for the endogeneity problem. Comparing these models in terms of sample-size retention (i.e., the external validity of a corrective method) and bias reduction (i.e., the statistical conclusion validity of a corrective method), we place King and Nielsen's conclusions in the context of a more comprehensive examination of PSA. We conclude by summarizing key issues in using PSA to correct for endogeneity, and we review core guidelines for the use of PSA when conducting observational studies.

Review of Key Propensity Score Methods

The Definition and Properties of Propensity Scores

Regression analysis of any type assumes that independent variables used in the regression are not correlated with the residual/error term. When researchers use

quasi-experimental designs, they typically assign study participants into treatment and comparison conditions based on nonrandom criteria. The presence of various selection effects can result in the residual term being correlated with the independent variables. Running a regression without controlling for these selection effects or biases results in inefficient and potentially misleading estimates of treatment effects. This is known as the *endogeneity problem*. In 1983, Rosenbaum and Rubin published a seminal article that proposed to use a propensity score to address the problem. The theories and application principles developed from their work laid the foundation for the entire class of propensity score models.

With complete data, Rosenbaum and Rubin (1983) defined the propensity score for participant i ($i = 1, \ldots N$) as the conditional probability of assignment to a particular treatment ($W_i = 1$) versus nontreatment ($W_i = 0$) given a vector of observed covariates, x_i: $e(x_i) = \text{pr}(W_i = 1 \,|\, X_i = x_i)$. The propensity score $e(x_i)$ is a balancing measure (called the coarsest score) that summarizes the information of vector x_i in which each x covariate is a finest score.

Several methods are available for estimating the conditional probability of receiving treatment using a vector of observed covariates. These methods include logistic regression, the probit model, and discriminant analysis. Of these methods, logistic regression is the prevailing approach. A closely related method uses the Mahalanobis metric distance. In these models—developed prior to PSM methods—Mahalanobis distance serves a similar function as a propensity score and is used as a basis for matching estimators. More recently, new models—such as the application of generalized boosted modeling (McCaffrey et al., 2004)—have been developed to refine logistic regression and, in turn, to refine propensity score estimation.

The greatest advantage of the propensity score is its reduction in dimensions, which solves the problem of an insufficient number of sample cases in exact matching. In practice, researchers must often correct for many covariates, which represent many dimensions. The propensity approach reduces these dimensions to a one-dimensional score. In conventional 1:1 matching, as the number of matching variables increases, the researcher is challenged by the difficulty of finding a good match from the control group for a given treated participant. Rosenbaum (2002) illustrated this with p covariates: Even if each covariate is a binary variable, there will be 2^p possible values. Suppose $p = 20$, then $2^{20} = 1{,}048{,}576$, or more than a million possible values of 20 covariates. With a sample of hundreds or thousands of participants, it is likely that many participants will have unique covariate values and, therefore, will be unmatchable. Matching in this context often results in dropping cases and, in the presence of a large number of covariates, may become infeasible. Rosenbaum and Rubin (1983) derived and proved a series of theorems and corollaries to justify three key approaches using the propensity scores: pair matching, propensity score subclassification, and covariance adjustment. These methods have been greatly expanded

since. Today, researchers may use at least eight closely related but technically distinct models. To frame our discussion of matching on propensity scores, we briefly review these PSA methods in the following section.

Propensity Score Methods

Nearest Neighbor Within Caliper Matching

This method aims to match each treated participant to one or more control participant based on estimated propensity scores. Denote P_i and P_j the propensity score for treated and control participants, respectively, I_1 the set of treated participants, and I_0 the set of control participants. A neighborhood, $C(P_i)$, contains a control participant j (i.e., $j \in I_0$) as a match for treated participant i (i.e., $i \in I_1$) if the absolute difference of propensity scores is the smallest among all possible pairs of propensity scores between i and j:

$$C(P_i) = \min_j |P_i - P_j|, j \in I_0.$$

Once j is found to match i, j is removed from I_0 without replacement. If for each i only a single j falls into $C(P_i)$, then the matching is nearest neighbor pair matching or one-to-one matching. If for each i there are n participants found to fall into $C(P_i)$, then the matching is 1-to-n matching.

In nearest neighbor matching, there is no restriction imposed on the distance between P_i and P_j as long as j is a nearest neighbor of i in terms of the estimated propensity score. By this definition, even if $|Pi - Pj|$ is large (i.e., j is very different from i on the estimated propensity score), j is still considered a match to i. To overcome shortcomings of erroneously choosing j, researchers must select j as a match for i only if the absolute distance of propensity scores between the two participants meets the following condition,

$$|P_i - P_j| < \varepsilon, j \in I_0,$$

where ε is a prespecified tolerance for matching, or a caliper. Rosenbaum and Rubin (1985) suggested using a caliper size of a quarter of a standard deviation of the sample estimated propensity scores (i.e., $\varepsilon \leq .25\sigma_P$, where σ_P is the standard deviation of the sample's estimated propensity scores).

Nearest neighbor matching within a caliper is a combination of the two approaches just described. Using these approaches, researchers begin by randomly ordering the treated and nontreated participants. They then select the first treated participant (i) and find j as a match for i if the absolute difference of propensity scores between i and j falls into a predetermined caliper (ε) and is the smallest among all pairs of absolute differences of propensity scores between i and other js within the caliper. Both i and j are then removed from consideration for matching, and the next treated participant is selected.

Mahalanobis Distance Matching

Mahalanobis distance matching (MDM) requires randomly ordering study participants and then calculating the distances between the first treated participant and all controls, where the distance—$d(i, j)$—between a treated participant (i) and a nontreated participant (j) is defined by the Mahalanobis metric distance: $d(i, j) = (u - v)^T C^{-1}(u - v)$, where u and v are values of the matching variables for treated participant i and nontreated participant j, respectively, and C is the sample variance–covariance matrix of the matching variables. The nontreated participant (j) with the minimum distance $d(i, j)$ is selected as the match for treated participant i, and both participants are removed from the pool. This process is repeated until matches are found for all treated participants. The covariates included in u, v, and C may or may not include an estimated propensity score. If these covariates include an estimated propensity score—$\hat{e}(x)$—then the method is called MDM with propensity scores; otherwise, it is called MDM without propensity scores.

Coarsened Exact Matching

Coarsened exact matching (CEM) is similar to exact matching but uses coarsened or less restrictive criteria (Iacus et al., 2011). The greatest advantage of this method is its ease. CEM does not use a one-dimensional score in matching, but like other propensity score models, it reduces categories of matching variables by using coarsened standards in such a way that continuous covariates are coarsened at natural breakpoints, such as high school and college degrees instead of years of education. Discrete variables are left as is or are combined, such as when researchers combine strong and weak Democrats into one category and strong and weak Republicans into another. Matching is done on these coarsened variables.

Optimal Matching

Both nearest neighbor within caliper matching and MDM are criticized for their requirement for a sizable common support region, which is defined as a region bounded by the maximum value of estimated propensity scores for the treated participants and by the minimum value of the estimated propensity scores for the nontreated participants. When the common support region is small or does not exist, both matching algorithms will fail.

To address these problems, Rosenbaum (2002) developed an optimal propensity matching approach using network flow theory from operations research. The treated participants are set A and the controls are set B, with $A \cap B = \emptyset$. The initial number of treated participants is $|A|$ and the number of controls is $|B|$, where $|\bullet|$ denotes the number of elements of a set. For each $a \in A$ and each $b \in B$, there is a distance, δ_{ab}, with $0 \leq \delta \leq \infty$. The distance measures the difference between a and b in terms of their observed covariates, such as their difference on propensity scores or Mahalanobis metrics. Matching is a process to develop S strata ($A_1, \ldots A_s; B_1, \ldots B_s$) consisting

of S nonempty, disjoint participants of A and S nonempty, disjoint subsets of B, so that

$$|A_s| \geq 1,\ |B_s| \geq 1,\ A_s \cap A_{s'} = \varnothing \text{ for } s \neq s',$$

$$B_s \cap B_{s'} = \varnothing \text{ for } s \neq s',$$

$$A_1 \cup A_2 \cup \ \ldots \cup A_s \in A, \text{ and}$$

$$B_1 \cup B_2 \cup \ \ldots \cup B_s \in B.$$

By this definition, a matching process produces S matched sets, each of which contains $|A_1|$ and $|B_1|$, $|A_2|$ and $|B_2|$, ... and $|A_S|$ and $|B_S|$. Note that by definition, within a stratum or matched set, treated participants are similar to controls in terms of propensity scores. Depending on the structure (i.e., the ratio of the number of treated participants to control participants within each stratum) the analyst imposes on matching, they can classify matching into the following three types:

1. Pair matching: Each treated participant matches to a single control, or a stratification of (A_1, ... A_s; B_1, ... B_s) in which $|A_s| = |B_s| = 1$ for each s.
2. Variable matching: Each treated participant matches to, for instance, at least one and at most four controls. Formally, this is a stratification whose ratio of $|A_s|:|B_s|$ varies.
3. Full matching: Each treated participant matches to one or more controls. Similarly, each control participant matches to one or more treated participants. Formally, this is a stratification of A_1, ... A_s and B_1, ... B_s in which the minimum of $|A_s|,\ |B_s| = 1$ for each s.

Optimal matching is the process of developing matched sets (A_1, ... A_s; B_1, ... B_s) with a size of α, β to minimize the total sample distance of propensity scores. Formally, optimal matching minimizes the total distance Δ defined as

$$\Delta = \sum_{s=1}^{S} \omega(|A_s|, |B_s|)\delta(A_s, B_s),$$

where $\omega(|A_s|, |B_s|)$ is a weight function and $\delta(A_s, B_s)$ is the difference between treated and control in terms of their observed covariates, such as their difference on propensity scores or Mahalanobis distances. There are three ways to define the weight function (see Rosenbaum, 2002).

The optimal matching method uses network flow theory to form matched sets that minimize the total distance. A primary feature of network flow is that it concerns the cost of using b for a as a match, where a cost is defined as the effect of the pair (a, b) on the total distance (Δ). A primary advantage of optimal matching, particularly full and variable matching, is that the original sample size is retained.

Postmatching outcome analysis for matched samples generated by optimal variable matching or optimal full matching can be performed by assessing the sample average treatment effect (ATE) using a Hodges–Lehmann aligned rank test (Hodges

& Lehmann, 1962). Outcome analysis for matched samples generated by the optimal pair matching needs to control for the clustering effect within multiply matched participants and may be performed as a regression of difference scores (Rubin, 1979).

Propensity Score Subclassification

The central idea of subclassification was developed by W. G. Cochran (1968) and formulated before the development of PSA. A subclassification algorithm using propensity scores balances data through five consecutive steps. First, sort the sample by estimated propensity scores in an ascending order. Second, divide the sample into K strata using quantiles (quintiles, deciles, or other) of the estimated propensity scores. Third, evaluate the treatment effect by calculating the mean difference of outcome and the variance of differences between treated and control participants within each stratum, or by running a multivariate analysis of outcomes within each stratum as one does for samples generated by a randomized experiment. Fourth, estimate the mean difference ATE for the whole sample (i.e., all K strata combined) through aggregating. Fifth, test whether the sample difference on outcome is statistically significant.

Let $0 = c_0 < c_1 < c_2 < \ldots < c_K = 1$ be boundary values. Let B_{ik} be the indicators defined as $B_{ik} = 1$ if $c_{k\text{-}1} < e(x_i) < c_k$, or $B_{ik} = 0$ otherwise, and

$$B_{ik} = 1 - \sum_{k=1}^{K-1} B_{ik},$$

where i is the index of observation ($i = 1, \ldots n_k$; n_k is the number of observations in stratum k), k is the index of the stratum ($k = 1, \ldots K$), and $e(x_i)$ is the propensity score for i. Now, the ATE within stratum k can be evaluated by applying the standard estimator to stratum k, or $\hat{\tau}_k = \bar{Y}_{k1} - \bar{Y}_{k0}$, where

$$\bar{Y}_{k\omega} \frac{1}{n_{k\omega}} \sum_{i:\, W_i = \omega} B_{ik} \times Y_i, n_{k\omega} = \sum_{i:\, W_i = \omega}^{n_i} B_{ik},$$

and $\omega = 1$ or 0, indicating treatment or control status. The condition under which the constant propensity score property holds is that K is sufficiently large and the differences $c_k - c_{k-1}$ are small.

The ATE for the whole sample is then estimated by using the weighted average of the within-stratum estimates:

$$\hat{\tau} = \sum_{k=1}^{K} \frac{n_k}{N} [\bar{Y}_{0k} - \bar{Y}_{1k}] \text{ for mean, and}$$

$$\hat{\tau} = \sum_{k=1}^{K} \frac{n_k}{N} [\hat{\tau}_k] \text{ for regression-type coefficient,}$$

where N is the total number of participants. The variance of the sample ATE is estimated by the following formulas:

$$Var(\hat{\tau}) = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 Var[\bar{Y}_{0k} - \bar{Y}_{1k}] \text{ for mean, and}$$

$$Var(\hat{\tau}) = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 Var[\hat{\tau}_k] \text{ for regression-type coefficient.}$$

Taking a square root of variance, one obtains a standard error of the ATE or regression-type coefficient and then can perform a significance test of a nondirectional or directional hypothesis.

Propensity Score Weighting

Propensity scores may be used without matching or subclassification in a fashion that is similar to data analysis with sampling weights. In this context, a weighted outcome analysis aims to increase internal validity, in a fashion similar to an analysis that uses sampling weights to increase external validity. Propensity score weighting is also called inverse probability of treatment weighting (or the IPTW estimator).

Propensity score weighting consists of the following three steps: First, estimate propensity scores using sample observed covariates (x) in a logistic regression or similar model. Next, calculate two types of weights: the weight for estimating ATE and the weight for estimating average treatment effect for the treated (ATT). For ATE, the weights are defined as follows:

$$\omega(W, x) = \frac{W}{\hat{e}(x)} + \frac{1 - W}{1 - \hat{e}(x)}.$$

By this definition, when $W = 1$ (i.e., a treated participant), the weight becomes

$$\omega(W, x) = 1/\hat{e}(x).$$

When $W = 0$ (i.e., a control), the weight becomes

$$\omega(W, x) = \frac{1}{1 - \hat{e}(x)}.$$

For ATT, the weights are defined as follows:

$$\omega(W, x) = W + (1 - W)\frac{\hat{e}(x)}{1 - \hat{e}(x)}.$$

By this definition, when $W = 1$ (i.e., a treated participant), the weight becomes $\omega(W, x) = 1$; when $W = 0$ (i.e., a control), the weight becomes

$$\omega(W, x) = \frac{\hat{e}(x)}{1 - \hat{e}(x)}.$$

Finally, specify the weight in an outcome analysis that treats the weight just like a sampling weight. The outcome analysis then becomes a propensity score weighted analysis. In the outcome analysis, if researchers also need to use sampling weight,

they can multiply the two types of weights (i.e., the sampling weight and the propensity score weight) and use the product weight in the outcome analysis (DuGoff et al., 2014).

Because propensity score weighting retains the original sample and does not trim a large number of cases, it is widely used in social behavioral and health research in conjunction with complicated outcome analyses, such as Cox proportional hazards and structural equation modeling.

Matching Estimators

Matching estimators refer to a collection of special matching algorithms developed by Alberto Abadie and Guido Imbens, including the simple matching estimator, the bias-corrected matching estimator, the variance estimator assuming homoscedasticity, and the variance estimator allowing for heteroscedasticity (Abadie et al., 2004; Abadie & Imbens, 2006). The crucial idea of this method is to impute the missing outcome or counterfactual at the unit level and then use both the observed and imputed values to evaluate a series of treatment effects.

Based on the counterfactual framework (Neyman, 1923/1935; Rubin, 1974, 1986), the matching estimators directly impute the missing data at the unit level by using a vector norm. That is, at the unit level, a matching estimator imputes potential outcomes for each study participant. Specifically, it estimates the value of $Y_i(0)|W_i = 1$ (i.e., potential outcome under the condition of control for a treatment participant) and the value of $Y_i(1)|W_i = 0$ (i.e., potential outcome under the condition of treatment for a control participant). After imputing the missing data, matching estimators can be used to estimate various ATEs, including the sample ATE, the population ATE, the sample ATE for the treated, the population ATE for the treated, the sample ATE for the controls, and the population ATE for the controls.

The matching estimators do not use logistic regression to predict propensity scores. Instead, these methods use a vector norm to calculate distances on the observed covariates between a treated case and each of its potential control cases, and distances on the observed covariates between a control case and each of its potential treated case. A minimum distance shows who is the match and determines the imputed counterfactual value. The *vector norm* uses the same formula as the Mahalanobis metric distance. Abadie and Imbens allow the use of both the sample variance–covariance matrix and the sample variance matrix (i.e., a diagonal matrix) in the computation, while the Mahalanobis method only uses the variance–covariance matrix. Abadie and colleagues (2004) provided a variance estimate for each of the six treatment effects, assuming homoscedasticity, so that analysts can use appropriate standard errors to perform significance tests. When treatment effects are heteroscedastic, Abadie et al. (2004) developed a robust estimator of variance, which involves a second matching procedure such that treated units are matched to treated units and control units are matched to control units. When one or more matching variables are continuous,

the bias-corrected estimator can be used to correct for biases due to inexact matching, which uses a least squares regression that adjusts the difference within the matches for the differences in their covariate values.

PSM With Nonparametric Regression

Propensity score analysis with nonparametric regression was developed by James Heckman, Hidehiko Ichimura, and Petra Todd (1997, 1998). A central feature of this method is the application of nonparametric regression (i.e., local linear regression with a tricube kernel, also known as *lowess*) to smooth unknown and possibly complicated functions. The method allows estimation of ATT by using information from all possible controls within a predetermined span. Because of this feature, the method is sometimes called kernel-based matching (Heckman et al., 1998). The model may be called a difference-in-differences approach (Heckman et al., 1997) when it is applied to data for two time points (i.e., pre- and posttreatment data) to show change triggered by an intervention in a dynamic fashion.

I_0 and I_1 are the indices for controls and treated participants, respectively, and Y_0 and Y_1 are the outcomes of control cases and treated cases, respectively; n_1 is the number of treated cases, and S_p is the common support region. To estimate a treatment effect for each treated case $i \in I_1$, outcome Y_{1i} is compared with an average of the outcomes Y_{0j} for matched case $j \in I_0$ in the untreated sample. Matches are constructed on the basis of propensity scores $\hat{e}(x)$ that are estimated using the logistic regression on covariates x. Precisely, when the estimated propensity score of an untreated control is closer to the treated case $i \in I_1$, the untreated case gets a higher weight when constructing the weighted average of the outcome:

$$\text{ATT} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_p} \left[Y_{1i} - \sum_{j \in I_0 \cap S_p} W(i,j) Y_{0j} \right],$$

where $W(i, j)$ is the weight estimated via *lowess*. When a data set including two time points is available, researchers can compute the change score on the outcome $(Y_{1ti} - Y_{1t'i})$ for the treated case i and outcome $(Y_{0tj} - Y_{0t'j})$ for the control case j, respectively. To conduct a difference-in-differences (DID) analysis, replace Y_{1i} with $(Y_{1ti} - Y_{1t'i})$ and Y_{0j} with $(Y_{0tj} - Y_{0t'j})$; the estimate is a special version of the ATT. The DID formula is

$$\text{DID} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_p} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_p} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

where $W(i, j)$ is the weight estimated via *lowess*. The application of *lowess* to matching is innovative. Because the asymptotic distribution of weighted averages is relatively complicated to program, no software packages currently offer parametric tests to discern whether a group difference is statistically significant. As a common

practice, researchers use bootstrapping to estimate the standard error of the sample mean difference between treated and nontreated groups.

In summary, over the past four decades, methods for program evaluation have undergone a significant change. Statisticians and econometricians have developed—and continue to develop—a range of propensity score and other models. The criticism and reformulation of the classical experimental approach symbolize a shift in evaluation. Although Rosenbaum and Rubin published their propensity score paper in 1983, debate about correction methods is lively today, and it has fueled the development of new approaches.

King and Nielsen's Critique and Comments

In 2016, Gary King and Richard Nielsen posted a working paper entitled *Why Propensity Scores Should Not be Used for Matching*, and the paper was published in 2019. They showed that the matching method often accomplishes the opposite of its intended goal—increasing imbalance, inefficiency, model dependence, and bias. To overcome the problems of matching, particularly those pertaining to nearest neighbor within caliper matching (NNWC), King and Nielsen recommended using MDM and coarsened exact matching (CEM) as alternatives. In the following section, we discuss King and Nielsen's paper, and we use it to motivate a renewed examination of the eight PSA models discussed earlier.

Focus on Classical Matching Methods

King and Nielsen criticized the use of propensity scores for matching, not the entire family of propensity score methods. They wrote, "We trace the PSM paradox to the particular way propensity scores interact with matching. Thus, our results do not necessarily implicate the many other productive uses of propensity scores" (King & Nielsen, 2019, p. 1). For those who use the terms PSA and PSM interchangeably, it is important to recognize that the current criticisms pertain only to matching.

Further, the title of King and Nielsen's paper is potentially misleading. The study criticized a very specific method of matching, that is, the classical NNWC model, not all matching methods. As we described earlier, other matching approaches exist, such as optimal matching, matching estimators, and PSM with nonparametric regression. Hence, by referring to NNWC as "matching," the title lacks specificity and precision. Potentially, it leads to misconception about the robustness, efficiency, and validity of other matching methods.

The Rosenbaum and Rubin Proof of the Properties of Propensity Scores

King and Nielsen criticized Rosenbaum and Rubin's proof regarding the properties of propensity scores. Although it was mathematically correct, they suggested that the proof is of little use and possibly misleading when applied to real data. They argued that the theorem encourages researchers to settle for the lower standard

of approximating only complete randomization and average levels of imbalance rather than a fully blocked randomized experiment, which has a higher likelihood of reducing model dependence. They also asserted that balancing only on a propensity score does not balance the entire vector of covariates: Equality between any two estimated scalar propensity scores does not imply that the two corresponding *k*-dimensional covariate vectors are exactly matched.

This critique ignores the fundamental property Rosenbaum and Rubin proved with regard to the propensity score: the reduction of dimensionality in matching. The endogeneity problem is synonymous with the violation of the strongly ignorable treatment assignment assumption: $(Y_0,Y_1) \perp W \mid X$, which states that conditional upon observed covariates X, the assignment of study participants to binary treatment conditions ($W = 1$ or $W = 0$) is independent of the outcome of nontreatment (Y_0) and the outcome of treatment (Y_1). Violation of this assumption leads to the call for using correction procedures such as a PSA model.

The key property of a propensity score is its summary of information of the entire set of covariates in X so that it becomes a scalar score. Rosenbaum and Rubin proved that after creating a propensity score, treatment assignment and the observed covariates are conditionally independent given the propensity score: $x_i \perp W_i \mid e(x_i)$. The conversion of $(Y_0,Y_1) \perp W \mid X$ into $x_i \perp W_i \mid e(x_i)$ underscores the importance and utility of the propensity score. That is, through the creation of a propensity score, multiple covariates are sufficiently reduced into one score. The key feature is that the balancing or coarsest score adequately summarizes the information of vector x_i, in which each x covariate is a finest score. The mathematical proof of this property is a major contribution made by Rosenbaum and Rubin.

After creating propensity scores, an analyst might still find that two participants with the same $\hat{e}(x_i)$ score have different values on some covariates in x. For instance, if gender is one covariate used in the estimation of propensity scores, two participants with the same value of $\hat{e}(x_i)$ still could be one female and one male. However, the mathematical proof shows that for the sample as a whole, the joint distribution of entire vector x is conditionally independent from the treatment assignment.

NNWC achieves a lesser degree of balance on covariates than MDM and CEM. The former approach does the matching by using a summary score through logistic regression, whereas the latter two methods perform matching based on individual covariates. The crucial task for all propensity score methods is to correctly specify the model estimating $e(x_i)$ to ensure the correct functional forms of the covariates used in the estimation; thus, the analyst can be sure that the scalar score correctly represents the joint distribution of all x covariates. The importance of ensuring covariate balance—and balance testing—is almost unanimously emphasized by PSA developers. It is routine to check covariate balance after running a corrective procedure and to rerun a model if there are imbalances on major covariates. The failure to achieve covariate balance should not be attributed to the use of propensity

scores in NNWC; rather, balance requires inclusion of appropriate predictors in the estimation model and development of the proper functional form after balance checks.

The Importance of Considering Both External and Internal Validities

An important issue pertaining to the advantages and disadvantages of NNWC, MDM, CEM, and all other PSA methods is the level of bias reduction (i.e., the statistical conclusion validity of a corrective method) and the extent to which each method retains the original sample size (i.e., the external validity of a corrective method). Both criteria should be used to evaluate the performance of a given approach.

For the three methods considered by King and Nielsen (2019), researchers often must choose between inexact matching and incomplete matching (Guo & Fraser, 2015; Parsons, 2001). While trying to maximize exact matches, the analyst may trim cases (i.e., drop unmatchable participants), which results in incomplete matching (i.e., the loss of cases and a reduction in sample size). Conversely, while trying to maximize cases, the analyst may create a sample with inexact matching results. By design, NNCM employs a summary balancing score and, therefore, trims fewer cases than MDM and CEM because the latter two methods match based on individual covariates. Hence, considering both external and internal validities of a correction method, NNCM is not always inferior to MDM and CEM.

In empirical research, researchers often prioritize retaining the original sample size because the sample represents the research population of interest. Whatever corrections researchers intend to make, the fundamental goal is to ensure that the resample or matched sample through propensity score modeling remains representative. In this context, the loss of a large proportion of sample observations is hazardous and should be avoided. The external validity of a correction method is a crucial criterion for evaluating the performance of the method.

Fully Blocked Randomization Versus Complete Randomization

King and Nielsen raised an important issue related to NNWC matching. NNWC aims only to approximate a completely randomized experimental design, whereas in most observational data sets a fully blocked randomized experimental design is preferred. Although this is an important comment, in practice, fully blocked designs require precise prior causal knowledge to select the correct blocking measure, precise measurement of the blocking variable, and exact matching. This is a tall order that is rarely accomplished in routine program evaluation. For instance, a widely used approach in health sciences and educational research is cluster randomization, in which groups of participating units (as opposed to individual units) are randomized. Cluster randomization without using covariates is not fully blocked, yet it is from this kind of randomization that we see the usefulness of approximating a

completely randomized experiment. Fully blocked designs are desirable but require blocking on a variable with high construct validity. The process of blocking typically produces inexact or incomplete matches.

Method to Compare Propensity Score Models

Monte Carlo Study

To illustrate the importance of considering both bias reduction and complete matching (sample-size retention) in the propensity score process, we conducted a Monte Carlo study to compare eight different models: OLS regression without correction of endogeneity, NNWC, MDM, optimal full matching, CEM, matching estimators, propensity score weighting, and propensity score subclassification. We compared all the models described earlier except the PSA with nonparametric regression, which we excluded because this approach estimates ATT and is not comparable with the other methods, which estimate ATE.

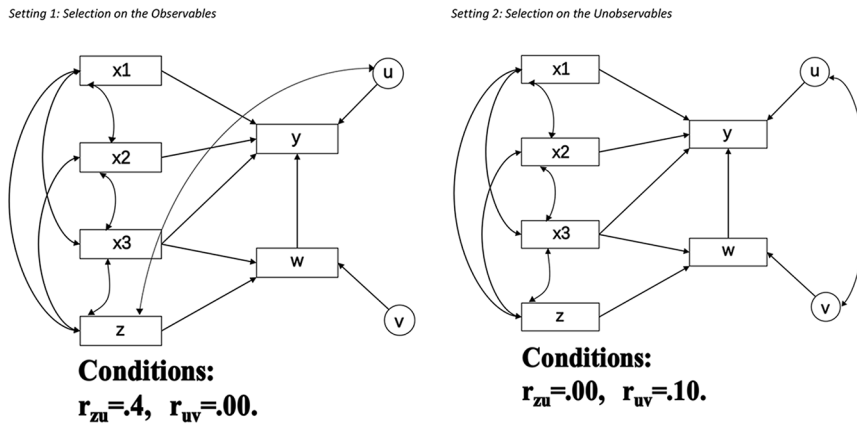
Following the design of a Monte Carlo study originally developed by Heckman and Robb (1985, 1986, 1988), the current study simulated two conditions of data generation: selection on the observables and selection on the unobservables. The former is defined by uncorrelated residuals in the sample selection equation and outcome equation. The latter is defined by the correlation of the two residuals. The data simulation compares the eight models based on both the level of bias reduction (i.e., the internal validity of the method) and mean observations retained by each model (i.e., the external validity of the method). Because one covariate used in the data simulation is continuous, the Monte Carlo study considered three categorizations of the continuous variable in CEM. Details of the Monte Carlo study specifications can be found in Guo and Fraser (2015), and the Stata syntax of the Monte Carlo study is available at the website for Guo and Fraser (2015). The current study contributes to the 2015 study by evaluating model performance on both sample-size retention and bias reduction. It also compares three new models (MDM with and without propensity scores, and CEM) recommended by King & Nielsen and one model (optimal matching) ignored by the 2015 study. The evaluation of model performance does not use the efficiency (or variance reduction) criterion because a comparison of model variance is complex and cannot be shown by data simulation. Properties of model efficiency are typically made analytically (see Guo & Fraser, 2015; Imbens & Wooldridge, 2009).

Findings

The design of the Monte Carlo study is shown in Figure 1, and results of the model comparisons under the two settings are shown in Tables 1 and 2.

Except for CEM with three categories, NNWC trimmed fewer cases than MDM and CEM, and this was true for both selection on observables and selection on

Figure 1. Design of the Monte Carlo Study: Two Settings of Selection Bias



Note. Sample size = 500; number of simulations = 10,000.

unobservables (Tables 1 and 2). When a continuous variable was categorized with a much-coarsened standard, its level of bias reduction was the worst among all models being compared, although it retains more observations. Considering both sample-size retention and bias reduction criteria, NNWC is not always inferior to MDM and CEM. Indeed, it is challenging to determine which of the three methods is best.

For selection on observables (Table 1), MDM with propensity score ranks Number 3, and MDM without propensity score ranks Number 1 in terms of bias reduction. Using this standard, MDM appears to be the preferred method when selection is measured. However, considering the sample size retained, MDM with propensity score ranks Number 11 and MDM without propensity score ranks Number 10, which are among the worse results across all models. Considering both internal and external validities, propensity score subclassification emerges as best.

For selection on unobservables (Table 2), neither MDM nor CEM can compete with the optimal full matching and matching estimators if both criteria are considered. Results also show that CEM in general is not robust in retaining observations and reducing bias, although the method has the advantage of easy application.

No single model works well across all scenarios. The "best" results depend on the fit between the assumptions embedded in a model and the process of data generation. For instance, under the setting of selection on observables (Table 1), because the data-generation process exactly meets the assumption embedded in the subclassification model (i.e., the number of strata is sufficiently large and the propensity score difference between two strata is small), the subclassifiction model ranks second in terms of bias reduction and appears to be preferable. However, for selection

Table 1
Results of Monte Carlo Study of Selection on the Observables

Analytic Model	Mean Observations Retained	Rank by Mean Observations	Mean Effect	Bias = Mean − True Effect	Rank by Mean Effect
Nearest neighbor within caliper	219	8	0.488	−0.012	5
Mahalanobis with propensity score	171	11	0.495	−0.005	3
Mahalanobis without propensity score	178	10	0.499	−0.001	1
Optimal full matching	500	1	0.395	−0.105	11
Coarsened exact matching (5 categories)	103	12	0.489	−0.011	4
Coarsened exact matching (4 categories)	184	9	0.488	−0.012	5
Coarsened exact matching (3 categories)	342	7	0.484	−0.016	7
Treatment effect model	500	1	1.929	1.429	12
Matching estimator	500	1	0.453	−0.047	10
Propensity score weighting (ATE)	500	1	0.484	−0.016	7
Propensity score stratification	500	1	0.497	−0.003	2
Ordinary least squares regression	500	1	0.537	0.037	9

on unobservables (Table 2), because of a nonzero correlation of residuals that makes the propensity score differences between strata large and nonignorable, the subclassification model ranks among the worst. These findings underscore the importance of understanding the assumptions related to each correction model and having deep, substantive knowledge sufficient to understand the risk of selection on unobserved variables. When information regarding the tenability of model assumptions is not available, findings must be conditioned on a discussion of model assumptions. Seeking concordance across findings from multiple models is indicated.

In general, our study confirms that the three methods recommended by Imbens and Wooldridge (2009)—propensity score subclassification, propensity score weighting, and matching estimators—are robust in most data situations.

Finally, this study supports a methodological caution made repeatedly by experienced observational researchers: OLS regression is a poor and ill-advised analytic

Table 2
Results of Monte Carlo Study of Selection on the Unobservables

Analytic Model	Mean Observations Retained	Rank by Mean Observations	Mean Effect	Bias = Mean − True Effect	Rank by Mean Effect
Nearest neighbor within caliper	218	8	0.646	0.146	5
Mahalanobis with propensity score	170	11	0.676	0.176	7
Mahalanobis without propensity score	178	10	0.677	0.177	8
Optimal full matching	500	1	0.422	−0.078	2
Coarsened exact matching (5 categories)	100	12	0.698	0.198	12
Coarsened exact matching (4 categories)	179	9	0.657	0.157	6
Coarsened exact matching (3 categories)	341	7	0.642	0.142	4
Treatment effect model	500	1	0.505	0.005	1
Matching estimator	500	1	0.639	0.139	3
Propensity score weighting (ATE)	500	1	0.686	0.186	9
Propensity score stratification	500	1	0.693	0.193	11
Ordinary least squares regression	500	1	0.691	0.191	10

approach in the presence of endogeneity or selection bias. Using the bias-reduction criterion, OLS regression ranks as Number 9 for selection on observables (Table 1) and Number 10 for selection on unobservables (Table 2).

Conclusion

As King and Nielsen pointed out, model specification in PSA is challenging. Our findings suggest that researchers need comprehensive knowledge of model assumptions and knowledge of plausible causal structure. From prior research, sources of selection bias must be understood. Substantive knowledge of plausible causal structure typically includes the theory of change of an intervention program being evaluated, which determines the covariates that should be used in the model predicting propensity scores and in the outcome analysis. Sample reduction after running a propensity score model is a key issue and should always be considered. Using both

bias reduction and sample-size retention criteria, MDM and CEM cannot be assumed to be good choices. Our findings suggest that it is of paramount importance to understand the assumptions of propensity score models and attend to potential violations of these assumptions. This requires both methodological and substantive knowledge.

Author Notes

Shenyang Guo, PhD, is a visiting professor in the Department of Sociology at Xi'an Jiaotong University and the Frank J. Bruno Distinguished Professor at the Brown School of Social Work, Washington University in St. Louis.

Mark Fraser, PhD, is a professor emeritus at the School of Social Work, University of North Carolina at Chapel Hill.

Qi Chen, MA, is a PhD candidate in the Department of Sociology at Xi'an Jiaotong University.

Correspondence regarding this article should be directed to Shenyang Guo, One Brookings Drive, Campus Box 1196, St. Louis, MO 63130 or via e-mail to s.guo@wustl.edu.

Acknowledgments

We thank Linyun Fu for her excellent work with the literature review for this study.

References

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata Journal*, *4*(3), 290–311. https://doi.org/10.1177/1536867X0400400307

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313. https://doi.org/10.2307/2528036

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, *49*(1), 284–303. https://doi.org/10.1111/1475-6773.12090

Gum, P. A., Thamilarasan, M., Watanabe, J., Blackstone, E. H., & Lauer, M. S. (2001). Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *Journal of the American Medical Association*, *286*(10), 1187–1194. https://doi.org/10.1001/jama.286.10.1187

Guo, S. (2015). Shaping social work science: what should quantitative researchers do? *Research on Social Work Practice*, *25*(3), 370–381. https://doi.org/10.1177/1049731514527517

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, *28*(4), 357–383. https://doi.org/10.1016/j.childyouth.2005.04.012

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed). SAGE Publications.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*(4), 605–654. https://doi.org/10.2307/2971733

Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, *65*(2), 261–294. https://doi.org/10.1111/1467-937X.00044

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156–245). Cambridge University Press.

Heckman, J. J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–113). Springer-Verlag.

Heckman, J. J., & Robb, R. (1988). The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatment on outcomes. In G. Duncan & G. Kalton (Eds.), *Panel surveys* (pp. 512–538). John Wiley.

Hodges, J., & Lehmann, E. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*, *33*(2), 482–497. https://doi.org/10.1214/aoms/1177704575

Iacus, S. M., King, G., Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345–361. https://doi.org/10.1198/jasa.2011.tm09599

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

Jones, A. S., D'Agostino, R. B., Gondolf, E. W., & Heckert, A. (2004). Assessing the effect of batterer program completion on reassault using propensity scores. *Journal of Interpersonal Violence*, *19*(9), 1002–1020. https://doi.org/10.1177/0886260504268005

King, G., & Nielsen, R. (2019). *Why propensity scores should not be used for matching*. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Morgan, S. L. (2001). Counterfactuals, causal effect, heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, *74*(4), 341–374. https://doi.org/10.2307/2673139

Neyman, J. S. (1935). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society, Series B*, *2*, 107–180. (Original work published in Polish in 1923.) https://doi.org/10.2307/2983637

Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*(4), 387–398. https://doi.org/10.1016/S0895-4356(00)00321-8

Parsons, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques (SAS SUGI paper 214–26). *Proceedings of the 26th annual SAS Users' Group International Conference*. SAS Institute. http://www2.sas.com/proceedings/sugi26/p214-26.pdf

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*(366), 318–328. https://doi.org/10.1080/01621459.1979.10482513

Rubin, D. B. (1986). Which ifs have causal answers? *Journal of the American Statistical Association*, *81*(396), 961–962. https://doi.org/10.1080/01621459.1986.10478355

Manuscript submitted: March 20, 2019
First revision submitted: June 2, 2019
Second revision submitted: July 14, 2019
Accepted: July 22, 2019
Electronically published: October 16, 2020