Received: 3 April 2019 | Revised: 17 June 2020 | Accepted: 22 June 2020

DOI: 10.1002/pst.2051

MAIN PAPER

WILEY

To use or not to use propensity score matching?

Jixian Wang

Celgene International Sarl, Boudry, Switzerland

Correspondence
Jixian Wang, Celgene International Sarl, Route de Perreux 1, Boudry, Basel, Switzerland.
Email: jixwang@celgene.com

Summary

Propensity score matching (PSM) has been widely used to reduce confounding biases in observational studies. Its properties for statistical inference have also been investigated and well documented. However, some recent publications showed concern of using PSM, especially on increasing postmatching covariate imbalance, leading to discussion on whether PSM should be used or not. We review empirical and theoretical evidence for and against its use in practice and revisit the property of equal percent bias reduction and adapt it to more practical situations, showing that PSM has some additional desirable properties. With a small simulation, we explore the impact of caliper width on biases due to mismatching in matched samples and due to the difference between matched and target populations and show some issue of PSM may be due to inadequate caliper selection. In summary, we argue that the right question should be when and how to use PSM rather than to use or not to use it and give suggestions accordingly.

KEYWORDS

causal inference, dose-exposure-response relationship, health technology assessment, modeling and simulation

1 | INTRODUCTION

Propensity score matching (PSM)[1,2] was proposed more than 30 years ago for causal inference to eliminate or reduce biases when treatments are not randomized. It has been widely used in observational studies, in particular, in post-marketing safety and outcome studies. PSM is increasingly being used in other areas of pharmaceutical statistics such as indirect comparison of treatments using multiple data sources. Compared with other matching approaches, PSM is much easier to use, especially when there are multiple potential confounders, as it matches on the propensity score (PS), that is, the probability of being treated as a function of covariates, rather than on the covariates themselves. The performance of PSM has been extensively examined via simulations,[3-8] as well as in theory.[9,10]

Matching based on PS aims at creating a quasi-randomization to reduce confounding biases rather than improving efficiency of statistical inference. Although efficiency may not always be a major concern, some other issues such as covariate imbalance are. Recently King et al[11] showed that even with a caliper to control mismatch, PSM did not always reduce covariate imbalance. When covariates are roughly balanced between groups, PSM might increase imbalance rather than reduce it. Other issues of PSM included loss of sample in terms of size and data points and increasing dependency on a correctly defined model. Consequently, they suggested that PSM should not be used in practice. Relevant to these issues, some application results[12,13] showed inconsistency between estimates by PSM and those from randomized experiments for some large studies, indicating potential biases in PSM estimators. PSM was also criticized as "There is no way to know in advance whether the method will work.[13]"

Indeed, PSM has two major potential flaws. First, achieving a balanced PS does not guarantee balanced prognostic factors between treatment groups. Second, pruning due to matching may be harmful if it is implemented inappropriately. The latter is a common issue to all matching methods, but it is more problematic for PSM.

Nevertheless, although PSM does not take prognostic effects into account, under certain technical conditions, it is of equal percent bias reduction (EPBR),[14-16] that is, the percent bias reduction in the treatment effect estimate is the same as the imbalance reduction in the matching variable. In the setting of Ref [16], it is equivalent to equal percent imbalance reduction in *all* covariates in the PS model. We critically examine the evidence for and against PSM in the literature. In addition, we show that PSM has useful practical properties. Although EPBR is a long-run property, we show that this may also hold in the short run (eg, for a specific study), as described in a conditional version of EPBR. In a small simulation study we examine the impact caliper on biases due to imbalance in matched subjects and due to the difference between matched and target populations, and show that although the commonly used caliper width of 0.2SD seemed a good choice to control the bias of imbalance within matched samples, it may not be adequate to control the population bias due to the difference between the matched and the target populations. In summary, we argue that there is no simple answer to whether PSM should or should not be used in practice. Instead, we make recommendations on when and how PSM should be used, including how to use it in combination with other approaches.

2 | PSM AND COVARIATE IMBALANCE

PSM is a combination of two powerful approaches: matching and PS, each of which have been used on its own in observational studies. The basic idea of matching is to match treated subjects with controls so that, within matched pairs, covariates values are the same or similar, hence treatment effects can be estimated by comparisons within matched pairs. It reduces the influence of these covariates, and consequently the variability and potential bias in matching based estimators if some of the covariates are confounders (ie, those affecting both the outcome and treatment allocation). However, often it is difficult, if not impossible, to match on multiple covariates. Instead of matching on them directly, PSM matches on the PS, a scalar, hence it is much easier than matching on multiple covariates, and yet it controls confounding biases, as matched subjects have similar possibility of being treated. Therefore, PSM creates a quasi-randomization environment so that a direct comparison between matched treated and control subjects can be carried out without the need of further adjustment.

However, the simple solution of PSM comes with costs. Ignoring prognostic effects of confounders not represented by the PS often leads to efficiency loss. In typical epidemiology studies where sample sizes are rather large, a loss of efficiency is not a major concern, hence a simple estimator independent of outcome models may be the best choice. For other purposes such as indirect comparisons between studies and/or data sources with small to moderate sample sizes, approaches with sufficient efficiency should be used, hence, a simple PSM approach may not be a good choice. In fact, there are more important issues than the loss of efficiency. Recently, King et al[11] critically examined PSM and showed that using PSM "increases imbalance, inefficiency, model dependence, research discretion, and statistical bias at some point in both real data and in data generated to meet the requirements of PSM theory." The typical scenario they considered is the use of 1:1 PSM matching without replacement to match treated subjects to a much larger set of control subjects. As matching can be considered as pruning subjects with large difference in the PS, they investigated the impact of pruning with calipers. Using the mean distance in the covariate space between the treated and the nearest control as the index of imbalance, they showed that when matching with calipers of decreasing width, although the balance of covariates initially improved, it eventually severely worsened.

There are different views from King et al., especially on the issue of imbalance, but most have not been published. (See, eg, Reference 17) To examine the critiques from practical aspect, an empirical evaluation in pharmacoepidemiology studies has been reported in Reference 18. The results showed that although increased imbalance was found after heavy pruning with decreasing caliper width, it was much less significant than that was shown in Reference 11.

We will also concentrate on the imbalance issue with 1:1 PSM and will consider its EPBR property and a more general one. If a PSM is EBPR, it improves the balance of all covariates by the same percentage on average, but EBPR is not a guarantee for a particular matching. In practice, when a covariate is originally near balanced, PSM is more likely to worsen its balance rather than to improve it, as the coefficient for a near balance covariate in the PS model is very small, consequently the PSM would not try to improve its balance. This is a similar situation as regression toward the mean, since any variation (due to matching on other covariates) from a balanced state can only worsen it.

Compared with EPBR, monotonic imbalance bounding (MIB)[19] is a more general property regarding the imbalance of covariates measured by a desirable distance, for example, the sample mean difference of each factor, before and after matching. A matching achieved MIB if the imbalance of each covariate is bounded by a value, which can be either an absolute one, or a fraction of the original imbalance. The former one can bound the bias induced by the imbalance at a certain level, if the impact of the covariate on the outcome is known, while the latter ensures balance improvement after matching for each analysis. The coarsened value exact matching[20] is a simple MIB approach. Although in general PSM is not MIB, when covariates with coarsened values (eg, categorized continuous variables) are used as categorical variables in the PS model, with a near zero caliper, a PSM becomes coarsened value exact matching, hence is also MIB. Intuitively, this property follows the fact that subjects within the same coarsened category have the same unique PS value, assuming the estimated coefficients in the PS model for different covariates are not exactly the same. A formal proof can be found in.[21] This property is independent of PS model specification and whether these covariates are confounders or not, but it suggests to "overfit" the PS model with key prognostic factors, even when they do not affect treatment allocation.

3 | CONDITIONAL EPBR

Although EPBR only promises long run bias reduction, with a linear discriminant function as the PS, the PSM also has additional properties to EPBR. Specifically, a PSM is EPBR conditional on specific matching for a specific study. That is, the conditional bias reduces by the same percentage as the imbalance reduction in matched PS values for a specific matching.

First we formally introduce EPBR developed in a series of papers by Rubin et al[14-16,22-24] and borrow their notation. Let $\mathbf{X}$ be p-covariates following ellipsoidal distributions, or a mixture of them, for treated and control populations with different means but proportional variance-covariance matrices. The outcome Y is a linear function of $\mathbf{X}$, $Y = \boldsymbol{\beta}^T\mathbf{X}$. The goal of PSM is to match each treated subject to a control with the same or similar PS value. Following the references above, we assume that matching is based on a PS defined as a linear discriminant function $Z = \boldsymbol{\alpha}^T\mathbf{X}$, and there are n treated subjects and N, $N >> n$, controls to be matched to the treated ones. Y can be written as

$$Y = \rho Z + W \tag{1}$$

where W is independent of Z. PSM is to match on Z to reduce the difference in Y between matched treated and control subjects. Let $\bar{Y}_t$, $\bar{Y}_{rc}$, and $\bar{Y}_{mc}$ be the means of Y among treated subjects, all controls and matched controls, respectively. $\bar{Z}_t$, $\bar{Z}_{rc}$, $\bar{Z}_{mc}$, $\bar{W}_t$, $\bar{W}_{rc}$, and $\bar{W}_{mc}$ denote their counterparts for Z and W, respectively. Let $\bar{\Delta}_{yr} = \bar{Y}_t - \bar{Y}_{rc}$ and $\bar{\Delta}_{ym} = \bar{Y}_t - \bar{Y}_{mc}$ be the unmatched and matched mean differences in Y between treated and control subjects, respectively. $\bar{\Delta}_{zr}$, $\bar{\Delta}_{zm}$, $\bar{\Delta}_{wr}$, and $\bar{\Delta}_{wm}$ denotes their counterparts for Z and W, respectives. Then $E(\bar{\Delta}_{yr})$ and $E(\bar{\Delta}_{ym})$ are the biases before and after matching. $\boldsymbol{\alpha}$ can be estimated by fitting a logistic model for treatment groups with $\mathbf{X}$ as covariates, or based on the mean and variance-covariance matrix of $\mathbf{X}$ in each group.[16,22] The goal of PSM is to eliminate or reduce the matched bias $E(\bar{\Delta}_{ym})$ via matching on Z. It has been shown that matching on Z is EPBR[16,22] in the sense that

$$\frac{E(\bar{\Delta}_{ym})}{E(\bar{\Delta}_{yr})} = \frac{E(\bar{\Delta}_{zm})}{E(\bar{\Delta}_{zr})}, \tag{2}$$

which means, with an EPBR matching if the imbalance in Z reduces from $E(\bar{\Delta}_{zr})$ to $E(\bar{\Delta}_{zm})$, the bias in matched Y reduces by the same percent. For example, if an EPBR matching method reduces the bias in $\mathbf{Z}_i$ to $E(\bar{\Delta}_{zm}) = 0.5E(\bar{\Delta}_{zr})$, the bias in matched Y also reduces to $E(\bar{\Delta}_{ym}) = 0.5E(\bar{\Delta}_{yr})$ as long as there is no zero-element in $\boldsymbol{\alpha}$. EPBR leads to equal percent average imbalance reduction in any covariate in $\mathbf{X}$. For example, with $\boldsymbol{\beta} = (1, 0, ..., 0)$, the left hand side of (2) is the average imbalance reduction of the first covariate in $\mathbf{X}$. Additional properties relating to EPBR, such as the maximum imbalance reduction can be derived when $\mathbf{X}$ follows the normal distribution.[15,22] Practical aspect of using EPBR including predicting the effects of PSM given parameters in multiple normal distribution for $\mathbf{X}$ and simulation results without the distributional assumptions can be found in Reference 23.

It should be emphasized that EPBR is a long-run property as (2) is a relationship between expectations. However, without additional assumptions, if a PSM is EPBR, it is also conditional EPBR given $\bar{\Delta}_{zr}$ and $\bar{\Delta}_{zm}$ in a specific matching. From (1) we can derive

$$\begin{aligned}\bar{\Delta}_{yr} &= \rho\bar{\Delta}_{zr} + \bar{\Delta}_{wr}\\ \bar{\Delta}_{ym} &= \rho\bar{\Delta}_{zm} + \bar{\Delta}_{wm}.\end{aligned} \tag{3}$$

Therefore, $E\left(\bar{\Delta}_{ym}|\bar{\Delta}_{zm}\right) = \rho\bar{\Delta}_{zm}$ and $E\left(\bar{\Delta}_{yr}|\bar{\Delta}_{zr}\right) = \rho\bar{\Delta}_{zr}$ which leads to a conditional version of EPBR:

Definition: A matching is conditional EPBR (CEBPR) if

$$\frac{E\left(\bar{\Delta}_{ym}|\bar{\Delta}_{zm}\right)}{E\left(\bar{\Delta}_{yr}|\bar{\Delta}_{zr}\right)} = \frac{\bar{\Delta}_{zm}}{\bar{\Delta}_{zr}}, \tag{4}$$

that is, conditional on a matching that reduces the imbalance in Z by $\bar{\Delta}_{zm}/\bar{\Delta}_{zr}$ percent for a specific matching, the bias in $\bar{\Delta}_{ym}$ reduces by the same percent. CEPBR is a stronger assurance than EPBR for the performance of PSM for a specific study. The conditional variance of $\bar{\Delta}_{ym}$ depends on W only

$$\mathrm{var}\left(\bar{\Delta}_{ym}|\bar{\Delta}_{zm}\right) = \mathrm{var}(\bar{\Delta}_{wm}) = 2\mathrm{var}(W)/n_m, \tag{5}$$

where n_m is the number of matched pairs, since Z and W are independent in (1). The above results only needs the decomposition of Y given in (1), not the distributional assumptions imposed in References 16, 24. Nevertheless, it is unclear under what alternative conditions the decomposition is also possible.

Furthermore, we can show that EPBR holds approximately for the bias estimate $\bar{\Delta}_{ym}$ as well. Specifically, borrowing the concept of a classification being probably approximately correct in machine learning research, we introduce the following definition:

Definition: A matching is conditional probably approximately EPBR (CPAEPBR) if

$$P\left(1-\delta < \frac{\bar{\Delta}_{ym}}{\bar{\Delta}_{yr}} \Big/ \frac{\bar{\Delta}_{zm}}{\bar{\Delta}_{zr}} < 1+\delta | \bar{\Delta}_{zr}, \bar{\Delta}_{zm}\right) > 1-\epsilon \tag{6}$$

conditional on $\bar{\Delta}_{zr}, \bar{\Delta}_{zm}$, with constants $1 > \epsilon \geq 0$ and $1 > \delta \geq 0$.

In words, a PSM (or any other matching on a linear score) is CPAEPBR if with probability of at least $1-\epsilon$, EPBR holds with accuracy of at least δ. Here, the bias refers to the empirical bias in $\bar{\Delta}_{yr}$ and $\bar{\Delta}_{ym}$. ϵ and δ depend on the distributions of Z and W, the sample size and the matching result. CPAEPBR says that if we have reduced the imbalance in Z by $\bar{\Delta}_{zm}/\bar{\Delta}_{zr}$, it is likely the empirical bias $\bar{\Delta}_{ym}$ also reduces by a similar percentage. The EPBR rule does not hold exactly, but a large departure (measured by δ) from it is an event of small probability of less than ϵ. In practice, CPAEPBR makes sense only when ϵ and δ are much smaller than their upper bounds. Under the conditions given in References 16, 24, PSM is also CPAEPBR when n is large. The proof is technically involved, hence omitted here, except the following outline. Note that, conditional on $\bar{\Delta}_{zm}$ and $\bar{\Delta}_{zr}$, applying the central limit theorem leads to asymptotic normality of $\bar{\Delta}_{ym}$ and $\bar{\Delta}_{yr}$. Assuming $E\left(\bar{\Delta}_{yr}\right) = B > 0$ and $n \to \infty$, the ratio in (6) can be bounded around 1 based on the asymptotic normal distributions. Alternatively, the Chebyshev's inequality also leads to intervals which contain $\bar{\Delta}_{ym}$ and $\bar{\Delta}_{yr}$ with large probability, hence, (6) may hold without relying on asymptotic normality. Furthermore, (6) also holds unconditionally, as the probability of $\bar{\Delta}_{zm}/\bar{\Delta}_{zr}$ departures substantially from their mean ratio reduces to zero when $n \to \infty$.

The following simulation illustrates the meaning of CEPBR and CPAEPBR in practice. In the simulation, $\mathbf{X}$ was generated from a multi-variate normal distribution with $P = 5$, and covariance matrix $0.7\mathbf{I} + 0.3\mathbf{u}^T\mathbf{u}$, where $\mathbf{I}$ is a unit diagonal matrix of dimension p and $\mathbf{u} = (1, 1, 1, 1, 1)$. The means of $\mathbf{X}$ are 0 and $\mathbf{u}$ for control and treated subjects, respectively. We did not explicitly specify α but with above setting $\boldsymbol{\alpha}$ is proportional to $\mathbf{u}$. We set $\boldsymbol{\beta} = (1, 1, 0, 0, 0)$ so that it is fairly different from $\boldsymbol{\alpha}$. For each simulation, 100 treated and 500 control subjects were generated. α was estimated

by the mean and variance-covariace matrix of each treatment group, as detailed in References 14, 16. Each treated was matched to a control using the genetic matching algorithm implemented in R-package Matching.[25]

Figure 1 shows the relationship between $\bar{\Delta}_{ym}/\bar{\Delta}_{yr}$ and $\bar{\Delta}_{zm}/\bar{\Delta}_{zr}$ in 100 simulated studies. There is a clear linear relationship showing approximate proportionality between them, except for a small number of outliers. Other simulation (results not presented here) showed that, the number of outliers reduces by increasing the sample size and/or the similarity between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In summary CPAEPDR suggests that pruning to improve PS imbalance is likely beneficial even when PS is very different from the prognostic effect $\boldsymbol{\beta}^T\mathbf{X}$. However, as shown in the next section, over-pruning may lead to harm rather than benefit.

CEPBR does not guarantee average imbalance reduction for all covariates. As stated in Section 2, some covariates may be more imbalance after PSM for a specific study. One may wonder whether the post-PSM imbalance would have a considerable impact on $\bar{\Delta}_{zm}$, and consequently on $\bar{\Delta}_{ym}$. CPAEPDR guarantees that the chance for this to happen is small if the sample size is large and $\boldsymbol{\alpha}^T\boldsymbol{\beta}$ is bounded from zero. This may serve as a theoretical support for the empirical finding that worsened imbalance was mild and not common.[18]

4 | BALANCING BENEFIT AND HARM OF PRUNING WITH CALIPERS

The selection of caliper width is a key factor for a good PSM, for which there have been some simulation studies evaluating the performance of PSM in relation to caliper width selection.[26,27] Here we present a small simulation study that focuses on the topic discussed here, that is, the benefit and harm of pruning in a scenario that PSM works well if a proper caliper is used. The same data generation mechanism as in the previous section was used. We decompose $\bar{\Delta}_{ym}$ into two components:

$$\begin{aligned} \text{Population bias}: &\quad \bar{Y}_t - \bar{Y}_{mt}, \\ \text{Matching bias}: &\quad \bar{Y}_{mt} - \bar{Y}_{mc}. \end{aligned} \tag{7}$$

The first represents the empirical population difference in Y between the original treated and matched populations, and the second the difference between treated and control subjects within matched population. We evaluate the two biases separately as one may be more important than the other in a specific situation. For example, when treatment heterogeneity is unlikely or the matched population is meaningful in practice, the population bias may not be important. Our goal is to explore how the two biases change with a range of caliper width from 0.01SD to 1.5SD. Figure 2 presents the means of two biases, empirical SE and the number of matched pairs based on 10 000 simulation runs. As expected, increasing caliper width increased the number of matched pairs and reduces the population bias. However, it also

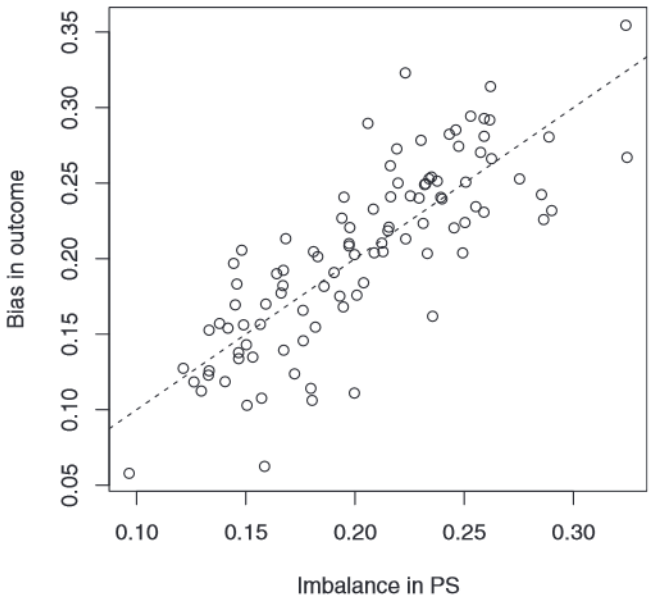

FIGURE 1 An example for conditional PAEPDR with 100 studies. Bias in outcome: $\bar{\Delta}_{ym}/\bar{\Delta}_{yr}$; imbalance in propensity score: $\bar{\Delta}_{zm}/\bar{\Delta}_{zr}$

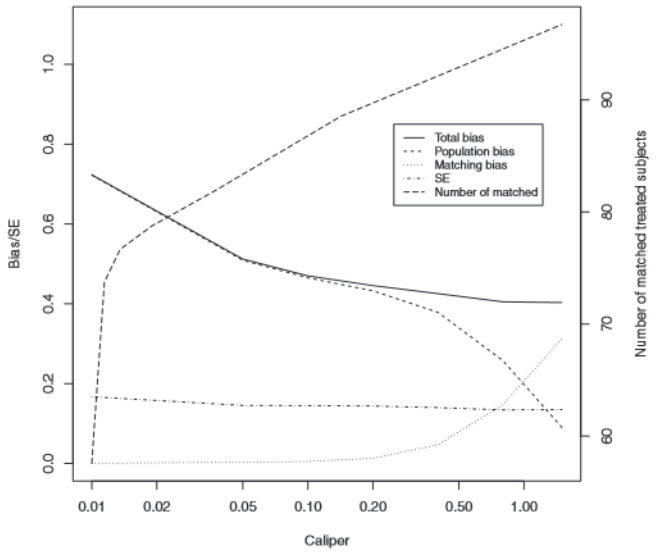

FIGURE 2 Empirical population, matching and total biases, SE and the number of matched treated subjects by caliper width (SD) with 10 000 simulation runs

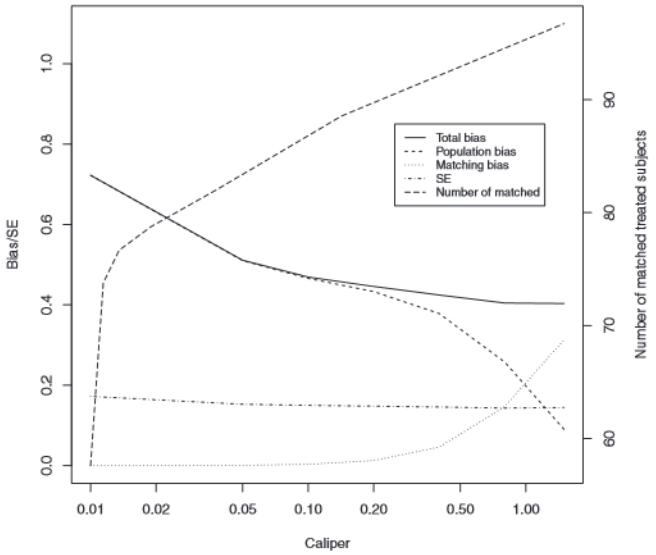

FIGURE 3 Estimated population, matching and total biases based on those in PS, SE in estimates and number of matched pairs by caliper width (SD) with 10 000 simulation runs

increased matching bias substantially when the caliper width was higher than 0.2SD. The result suggests that caliper width = 0.2SD seemed a good choice for matching bias control but may result in a large population bias. The empirical SE reduced slowly with the increase of caliper width. Note that population bias could be quite large even when there were only a few unmatched treated subjects, as they are likely to have covariate values very different from the matched ones. Therefore, one should carefully examine unmatched subjects and their impacts on the estimation.

In practice, one can also use the population and matching biases in the PS to evaluate the impact of caliper width upon treatment effect estimation based on $E\left(\bar{\Delta}_{ym}|\bar{\Delta}_{zm}\right)=\rho\bar{\Delta}_{zm}$. Figure 3 presents the same results as in Figure 2, but the biases were based on prediction using population and matching biases in the PS; the SE curve represents the mean estimated SE given by Matching package. The parameter ρ was estimated by regressing Y on Z. The results were very similar to those in Figure 2, suggesting that one can evaluate population and matching biases in the PS to assess the impact of caliper on PSM without repeating the analysis each time when a new matching is done with a different caliper width. This is particularly useful for planning the analysis when the PS and the PS-outcome relationship are available, but outcome data are not. The relationship between Y and Z may be nonlinear so EPBR may not hold, but evaluating population and matching biases in PS may still be useful to guide caliper selection.

5 | ALTERNATIVES TO PSM

In this section we consider alternatives to PSM in order to explore when they may work better, and how PSM can be modified or combined with an alternative in two directions. One is matching on alternatives to PS, another is using other approaches than matching.

Although matching on multiple covariates is often difficult, sometimes matching on individual coarsened values is feasible and may have more desirable properties than PSM due to MIB. One may also match on prognostic score (PGS), a measure of covariate contribution to the outcome, often determined by the control group.[28] In model (1) the PGS is $\boldsymbol{\beta}^T\mathbf{X}$. Matching on PGS is often more efficient as it is similar to the direct adjustment approach, and it can avoid some issues PSM has. However, it is difficult to implement because, to estimate PGS, outcome data have to be used and model selection may depend on them. Matching on PGS may also complicate statistical inference as it uses the outcome data twice, once for matching and once again for estimation. This issue is more problematic if the sample size is small. If feasible, matching on both PS and PGS is often a better approach than matching on only one of them, as it is doubly robust.[29] Matching on PS and one or more key prognostic factors has similar benefit, if feasible. The Mahalanobis distance, defined as the distance normalized by the variance-covariance matrix of the covariates, is yet another alternative to PS. The distance has been widely used in multi-variate analysis. However, matching on it may not be well justified, as the distance does not reflect the roles of covariates in either PS or PGS. In particular, using a caliper on Mahalanobis distance does not have the same direct control on imbalance in individual covariates as the coarsened value exact matching has.

PSM balances PS within matched pairs. Consequently, it also balances PS over the matched treated and control populations. In fact, to estimate causal treatment effect, the latter is sufficient and can be achieved in a different way, for example, by weighting based on PS or its alternatives. One of such approaches is the inverse probability of treatment weighting (IPTW) method, which weights each subject by the inverse probability of receiving the treatment actually received.[30] Therefore, for treated subjects, the inverse weights are the PS, while for the controls the inverse weights are one less the PS. Several papers have compared the performance of PSM and IPTW from either empirical or theoretical aspects. Hahn[31] derived semi-parametric lower variance bounds of IPTW estimator. Hirano et al.[32] showed that the bound can be achieved with estimated PS. In contrast, paired PSM does not reach the bound, although for 1: M matching the difference is bounded by $(2\,M)^{-1}$. However, IPTW requires a correctly specified PS model, for example, if the link function in the PS model is misspecified, the IPTW estimator would be invalid while the PSM estimator is rather robust to such misspecification. The doubly robust estimator,[33] which incorporates an outcome model into the IPTW estimator, is more robust as it is valid if either the outcome model or the PS models is correct, but it may perform badly when neither model is correct.

Balancing summary statistics such as the mean or proportion of covariates rather than PS over a population by weighting control subjects is also a popular approach, since it is much easier to achieve than matching on multiple covariates, although balance in the mean and proportion of key covariates does not guarantee equal joint distributions of them. This approach is in line with the common practice of checking covariate balance in observational studies and clinical trials. There is a large volume of literature on this topic, primarily in the area of survey sampling[34,35] and recently used for population matching in health economics evaluation.[36] It is also possible to (approximately) balance many covariates with proper bias control.[37] An obvious advantage of balancing covariates is that it is automatically MIB with a distance based on the sample mean or proportion. This approach can also be combined with IPTW to make the latter more robust by using the so-called covariate balancing PS (CBPS).[38] With CBPS, the weighted samples are balanced not only in the probability of being treated, but also in key covariates or functions of them. Therefore, it can be made multi-robust by balancing multiple potential prognostic scores. In fact, similar approaches exist in the toolkit for matching, known as matching with fine balancing.[39,40] These approach match subjects under the constraints of balancing some factors that are important but difficult to match.

6 | PRACTICAL SUGGESTIONS

Based on the properties and alternatives of PSM, together with the critiques of King et al., some suggestions on when and how to use PSM can be proposed. First of all, simple PSM such as greedy 1:1 matching without replacement,[2] which matches treated subjects in an arbitrary sequence and for each of them finds the best match among the unmatched controls, should be avoided unless for special cases. Often a better approach can be found that has most of

the advantages of a simple PSM but can also avoid the pitfalls King et al indicated. Typically, a better approach may be a PSM with simple modifications, or a combination with other approaches such as direct adjustment and balancing. The following are proposed suggestions in practical aspect:

1. Use PSM when matching on multiple covariates and/or direct adjustment with an outcome model are difficult or impossible. This may happen in drug safety studies in which the number of covariates is large, exposure-covariate data are rich, but safety events are rare;
2. Plan PSM carefully and evaluate its effect with available information at study planning stage;
3. Choose caliper width taking matching and population biases into account;
4. Evaluate imbalance in PS and key factors before and after matching, especially those originally well balanced;
5. Include prognostic factors into the PS model, even when they are far from statistically significant in the PS model. Machine learning approaches[41] can be used for variable/model selection, but key prognostic factors should be kept, regardless of its significance. The selected model should demonstrate acceptable balance in measured baseline variables between treated and control subjects;
6. Use categorical factors and coarsen covariates into categorical variables if appropriate;
7. Direct adjustment for key prognostic factors after PSM;
8. Match on both PS and PGS, if feasible. To avoid potential bias in the PSM based analysis due to PGS model selection, a possible remedy is to let independent statisticians to derive PGS independently from study statisticians;
9. Match on PS and key prognostic factors even they are well balanced before matching;
10. Directly or indirectly match key prognostic factors which may have a complex role in the PGS and balance the others.

Most of the suggestions are easy to implements. The list is by no means exhaustive, since other combinations of different approaches can be made for individual scenarios.

7 | DISCUSSION

After examining the above evidence, we can conclude that the right question should be when and how to use PSM. PSM is among the most useful tools for observational studies, if it is used properly, and often in combination with other approaches.

An issue not covered so far but worth mentioning concerns the interpretation of PSM estimators when a nonlinear outcome model is used. For example, if the outcome is binary then a logistic model stratified by prognostic factors provides the subject-specific log-odds ratio (log-OR), while an IPTW approach provides the population average one. However, the estimand corresponding to the estimator by stratification on matched pairs by PSM is not clearly defined. A recommendation for using PSM in survival analysis[42] applies here: use a nonstratified model to estimate, for example, the average treatment effect on the treated (defined in this scenario as the marginal log-OR in the treated population) and a stratified model to test (a sharp null hypothesis stronger than the one on the marginal log-OR).

It is important and feasible to plan PSM properly even before having access to full data. Under some rather ideal situations, the effects of PSM can be predicted based on statistics for covariates in both treatment groups.[14,15,22] The EBPR property and its conditional version suggest that assessing imbalance reduction in the PS is a useful way to evaluate the impact of bias reduction in causal effect estimates in the long run, and also for a specific study. The rather ideal situations hardly occur in practice, but the above approaches may serve as an approximation. Further work is needed to evaluate EBPR properties in less restrictive situations.

We have recommended, for PSM, to "overfit" the PS model with prognostic factors. When an IPTW approach is also used, for example, for sensitivity analysis, a natural question is whether the over-fitted PS model is also good for IPTW. The answer is positive, as overfitting a prognostic factor in the PS model also leads to more efficient IPTW estimators.[43]

The model (1) suggests a simple approach of including the PS as a covariate to directly adjust for confounding effects, also known as PS calibration.[44] This approach is sensitive to model misspecification, while PSM is rather robust to it. Therefore, although more difficult to implement, PSM is still a useful approach if robustness is important.

Although most of this article has been devoted to the discussion of method selection, one key recommendation is that, in practice, multiple methods should be applied, and results reported in detail and transparently. As some key assumptions cannot be verified by the data, the readers may review the results in their own way. One can specify a

primary analysis based on a specific method to address the issue of multiplicity, while the readers may have their own view on assumptions the primary analysis is based on. Overall, allowing the readers access to the results of multiple analyses would be mutually beneficial.

ACKNOWLEDGMENTS

The author would like to thank Drs A. Swern, M. Branson, M. Simcock, two referees, and the associate editor for very useful comments and suggestions which led to much improvement of this paper.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jixian Wang https://orcid.org/0000-0001-9963-6022

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
2. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The Am Stat*. 1985;39:33-38.
3. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51:171-184.
4. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29:2137-2148.
5. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Stat Med*. 2011;30:1292-1301.
6. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32:2837-2849.
7. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33:1057-1069.
8. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat Med*. 2014;33:4306-4319.
9. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006;74:235-267.
10. Abadie A, Imbens GW. Matching on the estimated propensity score. *Econometrica*. 2016;84:781-807.
11. King G, Nielsen R. Why propensity scores should not be used for matching. *Political Analysis*. 2019;27:435-454.
12. Smith JA, Todd PE. Reconciling conflicting evidence on the performance of propensity-score matching methods. *Am Economic Rev*. 2001;91:112-118.
13. Peikes DN, Moreno L, Orzol SM. Propensity score matching: a note of caution for evaluators of social programs. *The Am Stat*. 2008;62:222-231.
14. Rubin DB. Multivariate matching methods that are equal percent bias reducing, I: some examples. *Biometrics*. 1976;32:109-120.
15. Rubin DB. Multivariate matching methods that are equal percent bias reducing, II: maximums on bias reduction for fixed sampled sizes. *Biometrics*. 1976;32:121-132.
16. Rubin DB, Thomas N. Affinely invariant matching methods with ellipsoidal distributions. *Ann Statist*. 1992;20:1079-1093.
17. Jann B. 2018. Why propensity scores should be used for matching. Unpublished Presentation. https://boris.unibe.ch/101594/1/kmatch-kaiserslautern-2017.pdf. Accessed 13th July 2020.
18. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Implications of the propensity score matching paradox in pharmacoepidemiology. *Am J Epidemiol*. 2018;187:1951-1961.
19. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. *J Am Stat Assoc*. 2011;106:345-361.
20. Iacus S, King G, Porro G. Causal inference without balance checking: coarsened exact matching. *Polit Anal*. 2012;20:1-24.
21. Bestehorn F, Bestehorn M, Bestehorn M, Kirches C. A deterministic balancing score algorithm to avoid common pitfalls of propensity score matching. *arXiv*:1803.02704. 2019. Accessed 13th July 2020.
22. Rubin DB, Thomas N. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*. 1992;79:797-809.
23. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996;52:254-268.
24. Rubin DB, Stuart EA. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann Stat*. 2006;34:1814-1826.
25. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization. *J Stat Softw*. 2011;42:1-52.
26. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10:150-161.
27. Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *Am J Epidemiol*. 2014;179:226-235.

28. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95:481-488.
29. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med*. 2014;33:3488-3508.
30. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846-866.
31. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometnica*. 1998;66:315-331.
32. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71:1161-1189.
33. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Stat Sci*. 2008;22:523-580.
34. Deville JC, Särndal CE. Calibration estimators in survey sampling. *J Am Stat Assoc*. 1992;87:376-382.
35. Särndal CE. The calibration approach in survey theory and practice. *Surv Methodol*. 2007;33:99-119.
36. Phillippo D, Ades A, Sofia DS, Palmer S, Abrams K, Welton N. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2017;38:200-211.
37. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *J Am Stat Assoc*. 2015;110:910-922.
38. Imai K, Ratkovic M. Covariate balancing propensity score. *J Royal Stat Soc, Ser B*. 2014;76:243-263.
39. Rosenbaum PR. *Design of Observational Studies*. New York, NY: Springer; 2010.
40. Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J Am Stat Assoc*. 2015;110:115-127.
41. Cannas M, Arpino B. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biom J*. 2019;61:1-24.
42. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014;33:1242-1258.
43. Rotnitzky A, Li L, Li X. A note on overadjustment in inverse probability weighted estimation. *Biometrika*. 2010;97:997-1001.
44. Sturmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol*. 2005;162:279-289.

How to cite this article: Wang J. To use or not to use propensity score matching? *Pharmaceutical Statistics*. 2021;20:15–24. https://doi.org/10.1002/pst.2051