Wan *BMC Medical Research Methodology* (2025) 25:25
https://doi.org/10.1186/s12874-025-02481-w

BMC Medical Research
Methodology

RESEARCH Open Access

Propensity Score Matching: should we use it in designing observational studies?

Fei Wan[1]*

Abstract

Background Propensity Score Matching (PSM) stands as a widely embraced method in comparative effectiveness research. PSM crafts matched datasets, mimicking some attributes of randomized designs, from observational data. In a valid PSM design where all baseline confounders are measured and matched, the confounders would be balanced, allowing the treatment status to be considered as if it were randomly assigned. Nevertheless, recent research has unveiled a different facet of PSM, termed "the PSM paradox". As PSM approaches exact matching by progressively pruning matched sets in order of decreasing propensity score distance, it can paradoxically lead to greater covariate imbalance, heightened model dependence, and increased bias, contrary to its intended purpose.

Methods We used analytic formula, simulation, and literature to demonstrate that this paradox stems from the misuse of metrics for assessing chance imbalance and bias.

Results Firstly, matched pairs typically exhibit different covariate values despite having identical propensity scores. However, this disparity represents a "chance" difference and will average to zero over a large number of matched pairs. Common distance metrics cannot capture this "chance" nature in covariate imbalance, instead reflecting increasing variability in chance imbalance as units are pruned and the sample size diminishes. Secondly, the largest estimate among numerous fitted models, because of uncertainty among researchers over the correct model, was used to determine statistical bias. This cherry-picking procedure ignores the most significant benefit of matching design-reducing model dependence based on its robustness against model misspecification bias.

Conclusions We conclude that the PSM paradox is not a legitimate concern and should not stop researchers from using PSM designs.

Keywords Sample average treatment effect, Population average treatment effect, Imbalance, Model misspecification, Bias

Background

Propensity Score Matching (PSM) stands out as one of the most well-established and widely used strategies for exploring the comparative effectiveness of competing interventions in observational studies, such as comparing active treatment (referred to as "treated") to placebo control (referred to as "untreated") [1]. The propensity score (PS) represents the probability that a subject will receive the active treatment, given their baseline covariates [2]. Throughout this discussion, we assume that all baseline variables are measured, which is crucial for a valid PSM design. By capitalizing on the ignorability and balancing properties of PS, matching subjects based on their exact PSs ensures that the distribution of baseline variables becomes identical between treated and untreated individuals, and the treatment assignment for matched subjects can be

*Correspondence:
Fei Wan
wan.fei@wustl.edu
[1] Division of Public Health Sciences, Washington University in St Louis, 660 S. Euclid Ave, St Louis, MO 63110, USA

regarded as essentially random. PSM effectively enables us to extract an approximate randomized experiment from observational data, facilitating robust comparative analyses.

While numerous studies have meticulously examined and validated the properties of PSM through simulations and theoretical investigations [3–9], a more recent study by King and Nielson presents a contrasting view on PSM [10]. This study contends that PSM should not be used, as it paradoxically "increases imbalance, inefficiency, model dependence, research discretion, and statistical bias at some point in both real-world data and data tailored to adhere to PSM theory", a phenomenon they term the "PSM paradox". In the context of a one-to-one PSM design, the study illustrates that as study subjects were progressively pruned (e.g., by reducing the caliper size) and PSM was approaching exact matching, there was an initial improvement in balance. This progress continued until a specific point was reached, where covariates became nearly balanced, and PSM approximated a completely randomized design (CRD). However, it was at this juncture that the PSM paradox manifested itself, leading to a subsequent deterioration in balance, ultimately resulting in an increased bias in the effect estimation. Their findings have prompted researchers from various fields to cast widespread doubt on the validity of PSM [11–13].

Nevertheless, there are certain issues within their study that challenge the validity of their recommendation:

First, demonstrating the PSM paradox requires excessive pruning. After achieving initial balance with a reasonable caliper (observed confounding is minimized and a randomized design is mimicked), one must continue pruning instead of stopping at this point. Only then, using Mahalanobis distance and a drastically reduced sample size, does a jump in imbalance appear. It remains unclear why there is a need to further narrow the caliper width, potentially leading to the exclusion of valuable data, especially when baseline characteristics are already balanced using a reasonably sized caliper. For example, previous research has shown that using a caliper width of 0.2 times the standard deviation of logit PS can effectively eliminate over 90% of confounding bias [14].

Second, the study chose the sample average treatment effect or on sample effect on the treated (SATE or SATT) as the causal interest. While SATE or SATT represents the treatment effect for the specific study sample, it may diverge significantly from the population average treatment effect or population effect on the treated (PATE or PATT), especially when individual treatment effects exhibit heterogeneity, and the study sample is not randomly selected from the target population. Our primary interest often lies in estimating population-level quantities and the application of PSM for estimating population effects has been extensively investigated [3, 4].

Third, the study concludes that PSM mimics completely randomized designs (CRD), implying that every observation is independent. However, this contradicts previous recommendations advocating for matched-pair analyses in PSM designs [15]. Prior research [8] has demonstrated that exactly matched PSM designs can exhibit varying levels of within-pair intraclass correlation, indicating that PSM more closely resembles a randomized block design (RBD). The intraclass correlation quantifies the degree of similarity between individuals within the same block.

Fourth, the chosen imbalance metric, which calculates the average "pairwise" absolute distance in covariate space from each treated subject to the closest untreated unit, raises concerns. This metric's limitations are twofold: i) Even when PSs are identical, treated and untreated subjects in a matched pair typically exhibit covariate mismatches-a point already extensively discussed by Rosenbaum [16]. ii) These mismatches occur randomly, with negative and positive mismatches having a similar occurrence. As the number of matched pairs increases, the distributions of matched covariates eventually become similar between the treated and untreated groups, and the between-group imbalance becomes negligible, despite individual units within each matched pair having different covariate values. These are key implications of the balancing score and the strongly ignorability properties of PS. Common multivariate distance metrics, such as the Mahalanobis distance, may not fully capture the random nature of covariate imbalance in PSM designs. However, these metrics are valuable in covariate matching (CM) designs for assessing imbalance when matching is inexact, as this imbalance mainly contributes to residual confounding.

Lastly, Since the balancing score and the strongly ignorability properties of PS ensure that any mismatches in covariate values between matched pairs are random occurrences and don't lead to residual confounding, these mismatches can not bias the effect estimation. To demonstrate the increased statistical biases in the PSM paradox, the study adopts a biased approach for effect estimation, which involves the use of a well-known biased estimator: selecting the largest estimate from among hundreds of competing linear models. This rationale is based on the assumption that researchers may be uncertain about the correct model and should explore all possible models in post-matching analysis, potentially cherry-picking the best estimate. However, this approach neglects one of the most significant benefits of the matching design - its resilience against model misspecification. Importantly, this bias doesn't stem from imbalances in confounders between comparison groups.

In light of these uncertainties, our study examines whether the PSM paradox is a valid concern, whether the properties of PS apply in matching designs, and whether researchers should avoid using PSM in comparative effectiveness research. This study does not focus on comparing the advantages and disadvantages of PSM with other methods or designs.

Methods

Definitions and assumptions

The Rubin Causal Model (RCM), introduced by Rubin in 1974, is a widely used framework for defining causal effects [17]. In this model, we denote the binary treatment status as A, where 1 represents the treatment of interest, and 0 represents a control condition. Additionally, let $\mathbf{X}$ represent a $p \times 1$ vector of p confounders at baseline, and Y denote a continuous outcome variable. $Y(a)$ denote the potential outcome that would have been observed for an individual if the treatment A had been set to level a, where $a \in \{0, 1\}$. For simplicity, we assume that confounders $\mathbf{X} = \langle X_1, X_2, \cdots, X_p \rangle^T$ are mutually independent, i.e., $X_k \perp\!\!\!\perp X_j$ for all $k \neq j$.

We further make the following assumptions for causal inference:

Assumption 1 The Stable Unit Treatment Value Assumption (SUTVA), which consists of two sub-assumptions:

(i) The potential outcomes for any unit do not vary with the treatment assigned to other units (no interference between units).
(ii) For each unit, there are no different versions of each treatment level (no hidden versions of treatments).

Under SUTVA, the potential outcomes of each individual i depends only on the treatment assigned to this unit, not the treatments assigned to other units. For each individual i, $Y_i(1)$ represents the potential outcome that would have been observed if individual i received the treatment of interest, while $Y_i(0)$ represents the potential outcome if individual i received the control. The observed outcome is denoted as $Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$.

Assumption 2 The conditional ignorable treatment assignment assumption

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp A_i | \mathbf{X}_i$$

Conditional on observed confounders, the treatment status can be considered as randomly assigned.

We also impose the Positivity assumption as follows:

Assumption 3 For all observed covariates $\mathbf{X}$ where $P(\mathbf{X}) > 0, 0 < P(A = 1|\mathbf{X} = \mathbf{x}) < 1$

This is also known as the common support or overlap assumption because it entails that the conditional distributions $P(A = 1|\mathbf{X} = \mathbf{x})$ and $P(A = 0|\mathbf{X} = \mathbf{x})$ must share a common support.

Population and sample causal estimand

It's important to note that for any given individual i, only one potential outcome from the pair $\{Y_i(0), Y_i(1)\}$ can be observed. As a result, individual-level treatment effects $Y_i(1) - Y_i(0)$ cannot be identified, leading researchers to often focus on average treatment effects (ATE). In the literature, there are two types of ATEs: one at the population level and the other at the sample level.

We assume there is a population of N units in the super-population, from which we draw a sample S with n individuals. The population average treatment effect (PATE) is defined as

$$\begin{aligned}\tau_{PATE} &= \mathbb{E}(Y(1) - Y(0)) \\ &= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))\end{aligned}$$

which is the difference in potential outcomes averaged across the N units in the super-population. The sample average treatment effect (SATE) for our sample S is defined as

$$\begin{aligned}\tau_{SATE} &= \mathbb{E}(Y(1) - Y(0)|S = 1) \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\end{aligned}$$

which is the difference in the potential outcomes averaged across the n units in the sample S. SATE could vary from sample to sample if the individual treatment effect $Y_i(1) - Y_i(0)$ is not constant. If S is sampled randomly from the super-population, SATE is an unbiased estimate for PATE.

Similarly, we assume the number of treated subjects in the population and sample are N_1 and n_1. We can define the population and sample treatment effect among the treated as

$$\begin{aligned}\tau_{PATT} &= \mathbb{E}(Y(1) - Y(0)|A = 1) \\ &= \frac{1}{N_1}\sum_{i=1}^{N_1}A_i\big(Y_i(1) - Y_i(0)\big)\end{aligned}$$

$$\begin{aligned}\tau_{SATT} &= \mathbb{E}(Y(1) - Y(0)|A = 1, S = 1)\\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} A_i\big(Y_i(1) - Y_i(0)\big)\end{aligned}$$

PSM commonly targets the averaged treatment effect among the treated (ATT). Under the assumption that individual treatment effects are constant, i.e., $Y_i(1) - Y_i(0) = \tau$, $\tau_{PATE} = \tau_{SATE} = \tau_{PATT} = \tau_{SATT}$. Under SUTVA and the conditional ignorable treatment assignment assumption, we can identify PATE with observed outcome.

$$\begin{aligned}\tau_{PATE} &= \mathbb{E}(Y_i(1) - Y_i(0))\\ &= \mathbb{E}(\mathbb{E}(Y_i(1) - Y_i(0)|\mathbf{X}_i))\\ &= \mathbb{E}(\mathbb{E}(Y_i|A_i = 1, \mathbf{X}_i) - \mathbb{E}(Y_i|A_i = 0, \mathbf{X}_i))\end{aligned}$$

The relationship between Y, A, and $\mathbf{X}$ was previously described using the following linear model [10]:

$$Y_i = \mathbb{E}(Y_i|A_i, \mathbf{X}_i) + \epsilon_i = \beta_0 + \beta_1 A_i + g(\mathbf{X}_i) + \epsilon_i \quad (1)$$

where ϵ_i is a random error and $\mathbb{E}(\epsilon_i) = 0$, β_1 represents the conditional exposure effect. $g(\cdot)$ is some arbitrary function. When the effects of $\mathbf{X}_i$ on Y_i are linear additive, $g(\mathbf{X}_i) = \boldsymbol{\beta}_2^T\mathbf{X}_i$, where $\boldsymbol{\beta}_2$ represents the $p \times 1$ vector of regression coefficients for $\mathbf{X}$. It follows that $\beta_1 = Y_i(1) - Y_i(0)$ and $\tau_{PATT} = \tau_{SATE} = \tau_{PATT} = \tau_{SATT} = \beta_1$. King and Nielsen [10] stated that their causal interest lies in either SATE or SATT. However, model (1) implies a constant individual treatment effect. In this context, there is no distinction between population and sample causal estimands, nor between ATT and ATE. We adopted the same setting primarily to demonstrate that covariate imbalance in PSM occurs by chance and does not bias effect estimation, contrary to the findings of the prior study. We are not evaluating a novel estimation approach in more general settings. Finally, we focus on estimating the population effect, rather than the sample treatment effect, within our simulation design.

Propensity score

The PS, denoted as $e(\mathbf{X})$, is formally defined as the conditional probability of receiving an active treatment given the baseline covariates $\mathbf{X}$. It serves as a summary score of $\mathbf{X}$. The PS $e(\mathbf{X})$ has two key properties:

> **Property 1.** (**Balancing score**) The PS $e(\mathbf{X})$ balances the distribution of $\mathbf{X}$ between the treatment groups:
>
> $$A \perp\!\!\!\perp \mathbf{X}|e(\mathbf{X})$$
>
> When pairing two subjects, denoted as i and j, where one is treated and the other is untreated, such that $A_{ki} + A_{kj} = 1$, and they possess identical PSs in the kth matched pair, it's possible for these two subjects to exhibit different values for the observed covariates, $\mathbf{X}_{ki} \neq \mathbf{X}_{kj}$, despite having precisely the same PSs, $e(\mathbf{X}_{ki}) = e(\mathbf{X}_{kj})$. However, this discrepancy or mismatch in $\mathbf{X}$ within each matched pair can be attributed to chance and therefore cannot predict the treatment assignment [16]. The crucial aspect of these within-pair mismatches is that the disparity in covariate values between treated and untreated subjects can fluctuate randomly, resulting in both positive and negative differences from one matched pair to another, occurring with equal frequency. In situations where the number of matched pairs is small, we can anticipate moderate imbalance between the treated ($A = 1$) and untreated ($A = 0$) groups. However, as the number of matched pairs increases, the between-group imbalance becomes negligible. This distinctive phenomenon of PS can also be elucidated within the conventional regression framework [18]. Confounding bias emerges when confounders, which are correlated with both the outcome and treatment variables, are excluded from the regression model of Eq. (1). By adjusting for PS rather than directly including confounders as covariates, confounders are effectively decomposed into two components: the PS itself and a residual term. Conditioning on the PS, this residual term becomes orthogonal to the treatment variable and can be considered as random noise. Omitting such random noise no longer biases the estimation of treatment effect in a regression model.

It's important to note that successfully balancing the observed variables $\mathbf{X}$ by matching on $e(\mathbf{X})$ doesn't guarantee the balance of unmeasured variables. In practical applications, it is common to utilize the rule of thumb that considers a standardized mean difference (SMD), in absolute value, not exceeding 0.1 as a criterion to evaluate the balance of baseline variables individually between the treated and untreated groups. SMD of a confounder $X_j, j = 1, 2, \cdots, p$, is defined as:

$$d_j = \frac{\bar{X}_{1j} - \bar{X}_{0j}}{\sqrt{\frac{S_{1j}^2 + S_{0j}^2}{2}}}$$

where $\bar{X}_{1j}$, $\bar{X}_{0j}$, S_{1j}^2, and S_{0j}^2 are the sample means and sample variances of X_j in treated and untreated groups. SMD can capture both direction and size of the between-group imbalance in X_j.

Property 2. (**Strongly ignorable treatment assignment (SITA)**) If A is unconfounded given $\mathbf{X}$, then A is unconfounded given $e(\mathbf{X})$. Formally,

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A|\mathbf{X} \rightarrow \{Y(1), Y(0)\} \perp\!\!\!\perp A|e(\mathbf{X})$$

SITA requires that there are no unmeasured confounding variables and that there should be sufficient overlap in the PSs between the treated and untreated groups. If it suffices to match on $\mathbf{X}$ for a matching design, it also suffices to match on $e(\mathbf{X})$. Matching on a single variable $e(\mathbf{X})$ is more practical than matching on $\mathbf{X}$ when $\mathbf{X}$ is high-dimensional.

In summary, two properties of PS ensure that any imbalance in baseline confounders between two comparison groups in an exactly matched PSM design occurs by random chance and cannot result in residual confounding that can lead to biased estimation of treatment effect. Conversely, if covariate imbalances are systematic rather than random, property 2 will not hold due to residual confounding. It is important to note that a correctly specified PS model is required to estimate the PS unbiasedly. Otherwise, the estimated PS may not possess the balancing score and SITA properties, leading to biased results in subsequent methods, including PSM [19].

Finally, we assume $e(\mathbf{X})$ takes the following logit form:

$$e(\mathbf{X}) = P(A = 1|\mathbf{X}) = \frac{e^{\alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{X}}}{1 + e^{\alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{X}}} \tag{2}$$

where $\boldsymbol{\alpha_1}$ is the $p \times 1$ dimensional coefficient vector. We will use formula (2) in simulation study.

Results

The issues with the PSM paradox

Increase in imbalance

Instead of confirming the balancing property of the PS in matching designs, King and Nielsen [10] demonstrated that as PSM approaches exact matching by eliminating the worst-matched pairs, the imbalance initially decreases. However, it subsequently increases beyond a certain threshold, deviating from continual improvement. Since this observation contradicts the balancing property of the PS, we need to carefully examine whether the metrics used in previous studies can adequately capture the chance imbalance. King and Nielsen [10] used the following imbalance metric:

$$I(\mathbf{X}) = \text{mean}_{i \in \{i\}} d(\mathbf{X}_i, \mathbf{X}_{j(i)})$$

Here $I(\mathbf{X})$ represents the average pairwise distance between treated subject i with covariates $\boldsymbol{X_i}$, and the closest untreated subject j with covariates $\boldsymbol{X_{j(i)}}$. $d(\cdot)$ is a distance function. For example, the Mahalanobis distance is a popular choice, which is defined as the following:

$$d(\mathbf{X}_i, \mathbf{X}_{j(i)}) = \sqrt{(\mathbf{X}_i - \mathbf{X}_{j(i)})' \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_{j(i)})}$$

where Σ represents the sample covariance matrix of the original data. Ripollone et al. [20] found a similar pattern of increasing imbalance after progressive pruning of worst matched pairs in real epidemiological data sets. Instead of using the average individual distance between matched subjects, Ripollone et al. [20] measured imbalance using two different metrics: 1) the Mahalanobis distance between the covariate means in the treated and untreated groups, as follows:

$$d(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_0) = \sqrt{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)' \Sigma^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)}$$

Here $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_0$ are the vectors of covariate means in the treated and untreated groups. Larger values of the Mahalanobis balance metric suggest worse covariate balance. 2) C statistic for the discriminatory power of the logistic model predicting the treatment indicator in the matched data set. Higher C statistic values suggest worse covariate balance.

The balancing property of the PS suggests that as the caliper size shrinks and PSM approaches exact matching, the distribution of $\mathbf{X}$ becomes the same between the treated and untreated groups. Any mismatches in $\mathbf{X}$ between treated and untreated subjects in matched pairs are random occurrences. However, we need to understand that the balancing of $\mathbf{X}$ between the two groups is a large sample property. To better understand this concept, consider a small two arm CRD trial. In such cases, one might expect to observe significant mean differences in baseline covariates between treatment groups, even though treatments are randomly assigned. For instance, suppose a baseline variable X in the treated and untreated groups in a randomized trial follow a normal distribution $\sim N(\mu, \sigma)$. In this scenario, the group means of n subjects, denoted as $\bar{X}_1$ and $\bar{X}_0$, follow a normal distribution $\sim N(\mu, \frac{\sigma}{\sqrt{n}})$. SMD $d = \frac{\bar{X}_1 - \bar{X}_0}{\sigma} \sim N(0, \frac{2}{\sqrt{n}})$. When the sample size is small, the variance of d becomes large, making it more likely to observe a significant imbalance between two groups even in a randomized trial. Thus, in a PSM design with a large number of matched pairs, we would expect the between-group imbalance to be negligible when averaging out these within-pair mismatches. In repeated samples with a finite sample size (e.g., simulation studies), chance imbalance implies that the between-group imbalance can be positive in one matched sample and negative in another. However, when averaged across all matched samples, it converges to zero.

The issues with prior findings regarding the eventual increase in the between-group imbalance as PSM approaches exact matching are twofold:

(i) The distance metrics used in previous studies [10, 20] fail to capture the inherent "chance" aspect of observed imbalance in a PSM design. The Mahalanobis balance metric and *C* statistics, employed in these studies, measure the absolute distance and consistently yield positive values without considering the directions of either within-pair or between group imbalances. Consequently, when the number of matched pairs increases, the within-pair Mahalanobis distances [10] cannot average towards zero, and when averaged over repeated samples, the between-group Mahalanobis distances and *C* statistics [20] fail to converge towards zero as well.

(ii) Previous studies also demonstrated the PSM paradox by pruning the worst-matched pairs in a single dataset. However, balancing is a large sample property and the between-group imbalance in a PSM design tends to converge to zero with an increasing number of matched pairs, rather than a decreasing one from progressive pruning. When sample size is finite, this convergence towards zero occurs when averaged over repeated samples, not within a single sample. Instead, the noted rise in imbalance, as observed in prior studies [10, 20], reflects the growing variability in chance imbalance as the sample size decreases through progressive pruning. Once PSM reaches the initial balance of **X** with an appropriate caliper (i.e., the point where PSM approximates a randomized design), further reduction in matched pairs results in a smaller sample size, thereby increasing the likelihood of large chance imbalances between two groups. Consequently, this leads to large Mahalanobis distances or *C* statistics, as observed in previous studies [10, 20].This trend is akin to small trials where the likelihood of observing significant baseline covariate imbalances is higher.

Bias and model dependence

(i) *Bias*

As further revealed in the previous study [10], increasing imbalance has consequences that include a rise in bias. However, chance imbalance does not predict treatment status and should not bias the estimation of PATT when PSM approaches exact matching, even when using the sample mean difference, one of the simplest effect estimators [2]. So, where does this bias originate? It turns out that the bias observed in King and Nielsen's study [10] stems from an unconventional source - their choice of a generally biased estimator for the treatment effect. Consider a scenario in which an analyst has tried a set of different models $m_1, m_2, \cdots, m_J$ for estimating the treatment effect, resulting in corresponding estimates $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_J$ from each model. In such cases, researchers often opt for the maximum estimate among these, denoted as $\hat{\tau}_0 = \max(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_J)$. As stated in the previous study[10], this maximum coefficient $\hat{\tau}_0$ is typically biased, even when individual estimates are unbiased.

King and Nielsen [10] provided some reasons behind the use of this biased estimator in assessing PSM: The data generation process and the true model are unknown, which may lead analysts to explore various models, a practice commonly referred to as cherry-picking. In one of their examples, King and Nielsen fitted 512 different models for a simple PSM design with only two matching factors - an extreme case of post-matching analysis. Researchers often select the maximum estimate, which, according to order-statistics theory, is typically biased. Notably, this bias does not result from confounding, which arises from systematic imbalances in confounders.

(ii) *Model dependence*

King and Nielsen[10] also argue that human choices during cherry-picking in a PSM design can worsen model dependence. Previous works[10, 21] have attempted to define model dependence. Ho et al. [21] explained that the absence of model dependence means that the choice of a functional form does not significantly affect the results; the results remain consistent regardless of the selected functional form. King and Nielsen [10] further stated that the analysts encounter model dependence (i.e., different causal estimates from multiple equally well-fitting models) when exploring all plausible models. They proposed a formal metric to measure model dependence in matching designs by calculating the variance of effect estimates from all fitted models: $\hat{\sigma}^2 = \mathrm{var}(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_J)$[10]. This metric, based on the similarity of individual estimates from distinct effect estimators derived from competing models, primarily reflects a combination of design efficiency and the variance of each effect estimator, which inherently depends on the specific model fitted. However, it omits the critical bias component. This variance metric differs from the mean squared error (MSE) of a single estimator, which measures the average squared difference between estimated values and the true value, combining both variance and bias components for this estimator. Note that even two unbiased models can produce different individual estimates, depending on their variance estimators and sample size. We will use the following examples to illustrate the potential issues with this variance metric:

Example 1: Consider two competing designs, A and B, where all individual regression models produce unbiased estimates: $\hat{\tau}_{a1}, \hat{\tau}_{a2}, \ldots, \hat{\tau}_{aJ}$ and $\hat{\tau}_{b1}, \hat{\tau}_{b2}, \ldots, \hat{\tau}_{bJ}$, and $\mathbb{E}(\hat{\tau}_{aj}) = \mathbb{E}(\hat{\tau}_{bj}) = \tau, \forall j = 1, 2, \cdots, J$. Thus, unbiased estimation in both designs is independent of model specification, and any misspecified model can still yield unbiased estimates. Furthermore, let $\hat{\tau}_{a0} = \max(\hat{\tau}_{a1}, \hat{\tau}_{a2}, \ldots, \hat{\tau}_{aJ})$ and $\hat{\tau}_{b0} = \max(\hat{\tau}_{b1}, \hat{\tau}_{b2}, \ldots, \hat{\tau}_{bJ})$, $\hat{\sigma}_a^2 = \mathrm{var}(\hat{\tau}_{a1}, \hat{\tau}_{a2}, \ldots, \hat{\tau}_{aJ})$, and $\hat{\sigma}_b^2 = \mathrm{var}(\hat{\tau}_{b1}, \hat{\tau}_{b2}, \ldots, \hat{\tau}_{bJ})$. If $\hat{\tau}_{a0} < \hat{\tau}_{b0}$ and $\hat{\sigma}_a^2 < \hat{\sigma}_b^2$, can we then reverse the previous conclusion and claim that design A is less model-dependent than design B? Certainly not, as we know that unbiased effect estimation does not rely on model specification in either design.

Example 2: Consider two competing designs, A and B, where individual regression models produce unbiased estimates in design A but in design B only the correctly specified model yields unbiased estimates: $\mathbb{E}(\hat{\tau}_{aj}) = \tau, \forall j = 1, 2, \cdots, J$, and $\mathbb{E}(\hat{\tau}_{bj}) = \tau$ for $j = k$, while $\mathbb{E}(\hat{\tau}_{bj}) \neq \tau, \forall j \neq k$. If $\hat{\sigma}_a^2 = \hat{\sigma}_b^2$, can we conclude that both designs have the same level of model dependence? Certainly not reasonable, as we know unbiased effect estimation in design B depends on correct model specification.

In a randomized study, both CRD and RBDs allow unbiased estimation of treatment effects using sample mean differences without modeling the outcome-covariate relationship. With appropriate variance estimators, valid statistical inference can be achieved in either design, though the variance estimator in CRDs may be larger than in RBDs. However, lower efficiency does not compromise the validity of a design. In an observational study setting, both exactly matched PSM and CM designs can rely on the sample mean difference for unbiased effect estimation, as both designs effectively balance confounders. Similarly, with correct variance estimators, valid inference can be made in either design [22], without relying on a correctly specified outcome model. Although PSM could be less efficient than other CM designs given the same number of matched pairs, as it matches on a summary score rather than directly on covariates, using less information. However, it provides a practical solution to the curse of dimensionality, the key limitation of CM designs.

We propose that model dependence in matching designs is more appropriately defined by whether unbiased effect estimation relies on correct model specifications, as this property is a defining advantage of good matching designs. It is important to note that not all matching designs can reduce model dependence, such as matched case-control designs [23–25]. Using cherry-picking approaches to quantify bias or model dependence in randomized designs, or in valid matching designs that emulate randomization, is both invalid and unnecessary.

(iii) ***A good matching design eliminates the need for cherry-picking analysis and reduces the dependence of unbiased effect estimation on model specification***

Regression analysis with confounders adjusted as covariates in the original unmatched data is commonly used by applied researchers to estimate the treatment effect due to its relative simplicity compared to a matching design. The primary issue associated with regression analysis is model dependence, where the outcome model must be fitted correctly (e.g., considering nonlinear forms of confounders and interaction terms). A misspecified model can lead to biased estimates. In contrast, when using matching design, one must navigate a complex matching algorithm and may have to discard valuable unmatched observations. If, after all these painstaking efforts, researchers continue to grapple with the challenge of selecting the correct outcome model amidst the exploration of numerous candidate models during post-matching analysis, they should question the benefit of using matching design over regression adjustment in original data.

Researchers should note that matching in a cohort study can serve as a nonparametric preprocessing tool. When exact matching is achieved, it can either eliminate or reduce reliance on correct model specifications in the post-matching analysis, as discussed by Ho et al. [21]. In cases of exact matching, a simple linear regression that includes only the treatment indicator, essentially representing the sample mean difference between the treatment and control groups, or a model adjusted for covariates, can both yield unbiased estimates. This is true even if these models happen to be misspecified [26]. In situations where exact matching is not possible, the required adjustments are far less burdensome, less reliant on specific model assumptions than they would without matching. These informal claims find theoretical support in the work of Guo and Rothenhäusler [26]. They also demonstrated that when exact matching is unattainable, additional linear regression adjustments become necessary. Nonetheless, matching makes parametric analyses less sensitive to the correct model specification. These findings align with the earlier assertion that a well-designed matching process, coupled with subsequent regression adjustment, generally yields the least biased estimates [27]. Instead of experimenting with numerous models for post-matching analysis, we can use a predefined approach with commonly used misspecified models, such as simple linear regression or multiple linear regression including linear terms for all matching factors. The ultimate goal of comparative effectiveness research is unbiased effect estimation, not selecting results that appear most favorable. In the next section, we will assess the unbiasedness of these misspecified models using simulation.

The type of randomized design PSM emulates

PSM is suggested to mimic a CRD, whereas CM aims to replicate the randomized complete block design (RCBD) [10]. In general, CRD is less efficient than RBD. However, we show that PSM, rather than mimicking a CRD, emulates a form of RBD, where blocks are partial information of matched confounders $e(\mathbf{X})$, and the relative efficiency of exactly matched PSM versus CM also depends on the number of matched pairs.

For a 1:1 exact CM on confounders $\mathbf{X}$, M matched pairs are formed on $\mathbf{X}_m$, $m = 1, 2, \cdots, M$. The variances of the sample means $\bar{Y}_{i.} = \frac{\sum_{m=1}^{M} Y_{im}}{M}$, for $i = 0, 1$, and the mean difference $\bar{Y}_{1.} - \bar{Y}_{0.}$ are:

$$\text{Var}(\bar{Y}_{1.}) = \text{Var}(\bar{Y}_{0.}) = \frac{\boldsymbol{\beta_2}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta_2} + \sigma_\epsilon^2}{M},$$

and

$$\text{Var}(\bar{Y}_{1.} - \bar{Y}_{0.}) = \frac{2\sigma_\epsilon^2}{M}$$

Here, $\Sigma_{\mathbf{X}}$ denotes the $p \times p$ variance-covariance matrix of $\mathbf{X}$ in the matched population under CM. The intraclass correlation within pairs is:

$$\rho_1 = \frac{\boldsymbol{\beta_2}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta_2}}{\boldsymbol{\beta_2}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta_2} + \sigma_\epsilon^2} = \frac{1}{1 + \sigma_\epsilon^2 / \boldsymbol{\beta_2}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta_2}}$$

For a 1:1 matching design on exact logit PS, N matched pairs are formed on PSs Z_n, $n = 1, 2, \cdots, N$. While $\mathbf{X}_{1n} \neq \mathbf{X}_{0n}$, their PSs satisfy $e(\mathbf{X}_{1n}) = e(\mathbf{X}_{0n})$. The variance of the sample mean difference $\bar{Y}_{1.} - \bar{Y}_{0.}$ in this design is:

$$\text{Var}(\bar{Y}_{1.} - \bar{Y}_{0.}) = \frac{2(\sigma_\nu^2 + \sigma_\epsilon^2)}{N}$$

where $\sigma_\nu^2 = \sin(\theta)^2 \boldsymbol{\beta_2}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\beta_2}$ and $\tilde{\Sigma}_{\mathbf{X}}$ represent the $p \times p$ variance-covariance matrix of $\mathbf{X}$ in the matched population under PSM. Since $Z = \alpha_0 + \boldsymbol{\alpha_1}^T \mathbf{X}$, the intraclass correlation within pair is

$$\rho_2 = \frac{\gamma^2 \boldsymbol{\alpha_1}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\alpha_1}}{\gamma^2 \boldsymbol{\alpha_1}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\alpha_1} + \sigma_\nu^2 + \sigma_\epsilon^2} = \frac{1}{1 + \left(\sin(\theta)^2 \boldsymbol{\beta_2}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\beta_2} + \sigma_\epsilon^2\right) / \left(\left(\frac{\|\boldsymbol{\beta_2}\|}{\|\boldsymbol{\alpha_1}\|} \cos(\theta)\right)^2 \boldsymbol{\alpha_1}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\alpha_1}\right)}$$

Here θ is the angle between two coefficient vectors $\boldsymbol{\beta_2}$ and $\boldsymbol{\alpha_1}$. The sine distance, $\sin(\theta)$, measures the relationship between these vectors. When $\theta = 0$ or $\theta = \pi$, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\beta_2}$ are parallel, and the sine distance is minimized at 0. In this case, ρ_2 reaches its maximum:

$$\rho_{2,max} = \frac{1}{1 + \sigma_\epsilon^2 / \left(\left(\frac{\|\boldsymbol{\beta_2}\|}{\|\boldsymbol{\alpha_1}\|}\right)^2 \boldsymbol{\alpha_1}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\alpha_1}\right)}$$

When $\theta = \pi/2$, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\beta_2}$ are perpendicular, and the sine distance is maximized at 1. Here, ρ_2 reaches its minimum:

$$\rho_{2,min} = \frac{1}{1 + \left(\boldsymbol{\beta_2}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\beta_2} + \sigma_\epsilon^2\right)}$$

As θ increases from 0 to $\pi/2$ (or equivalently, from π to $\pi/2$), ρ_2 decreases from $\rho_{2,\text{max}}$ to $\rho_{2,\text{min}}$, transitioning from a RBD with a strong intraclass correlation to one with weak correlation. All the details are provided in supplemental material (Web Appendix 1).

We can make the following conclusions regarding relative efficiency of PSM compared to CM and the type of randomized design it mimics:

(1) PSM generally mimics a RBD. The intraclass correlation within pairs depends on the *sine* distance between $\boldsymbol{\beta_2}$ and $\boldsymbol{\alpha_1}$, with smaller distances yielding higher correlations. When the distance is large (i.e., $\sin(\theta) = 1$) and $\boldsymbol{\beta_2}^T \tilde{\Sigma}_{\mathbf{X}} \boldsymbol{\beta_2} + \sigma_\epsilon^2$ is substantial, ρ_2 becomes very small, making PSM approximate a CRD. However, upon closer examination of the settings in the previous study [10], the effects of confounders on the outcome and treatment appear parallel. Consequently, PSM essentially mimics a RBD with a high intraclass correlation within pairs ($\rho_2 > 0.7$), rather than a CRD, which would exhibit nearly zero intraclass correlation. Further details will be provided in the next section.

(2) The relative efficiency of the treatment effect estimator $\bar{Y}_{1.} - \bar{Y}_{0.}$ under PSM compared to CM can be expressed as:

$$\text{relative efficiency} = \frac{\sigma_\epsilon^2}{(\sigma_\nu^2 + \sigma_\epsilon^2)} \frac{N}{M}$$

If $N \approxeq M$, $\frac{\sigma_\epsilon^2}{\sigma_\nu^2+\sigma_\epsilon^2} < 1$, indicating that PSM is less efficient. This can be explained by the fact that PSM only use partial information of $\mathbf{X}$-"$e(\mathbf{X})$" as the blocking factor. However, when the number of matching variables is large, CM faces the curse of dimensionality, resulting in significantly fewer matched pairs compared to PSM ($M \ll N$). In such cases, the variance gains from CM may not offset the data loss, making PSM more efficient.

Simulation

Simulation design

We conducted simulation studies to achieve the following objectives: (a) *Assessing Systematic Imbalance*: Our first inquiry aimed to determine whether there exists a systematic imbalance and a corresponding bias in effect estimation as we tightened the matching caliper size. (b) *Evaluating Model Misspecification Sensitivity*: Our second inquiry focused on assessing whether PSM reduces sensitivity to model misspecification. (c) *Re-assessing the previous simulation study*: we will use the same data generation scheme used in previous simulation study [10] to show that PSM actually mimic a RBD in previous study, not a CRD.

In order to generate the simulation data, we followed the methodology used in our previous studies [8, 28], as detailed below:

(i) We created two coefficient vectors, $\boldsymbol{\beta}_2$ and $\boldsymbol{\alpha}_1$, each containing five elements in the outcome and treatment models (Eqs. (1) and (2)). For $\boldsymbol{\beta}_2$, we initiated the elements of the coefficient vector by randomly sampling values from the range of 1 to 9. Subsequently, we normalized this coefficient vector to be a unit vector. The sign of each element was determined using a Bernoulli distribution with a probability of 0.5. Finally, we set $\boldsymbol{\beta}_2$ equal to k multiplied by its normalized factor, with the value of k being fixed at 1.2. The same procedure was then repeated to generate $\boldsymbol{\alpha}_1$, with $\boldsymbol{\alpha}_1$ set to 1 multiplied by the normalized vector. Among all generated pairs of coefficients, we specifically selected two pairs of coefficients with their the *sine* distances falling within the intervals [0, 0.2] and (0.8, 1] respectively (Details in supplemental table 1 in Web Appendix 3). The *sine* distance measures the dissimilarity between two vectors, and it ranges from 0 to 1, with a larger value indicating a greater dissimilarity between the vectors [18]. When the *sine* distance falls within the range of (0.8, 1], it signifies that there is a weak within-pair correlation among the matched subjects. In this context, the PSM design closely resembles a CRD. Conversely, when the *sine* distance is within the range of [0, 0.2], it suggests that the within-pair correlation among matched subjects is strong. In such instances, the PSM design approximates a RBD. Our approach intends to avoid the extreme results from using proportional coefficient sets of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_2$ in prior studies [29].

(ii) For every pair of $\boldsymbol{\beta}_2$ and $\boldsymbol{\alpha}_1$, we generated five independent confounding variables, denoted as $X_1, X_2, \ldots, X_5$, from a normal distribution with mean 0 and standard deviation 1, each with a sample size of $n = 1500$. The treatment variable A was created using the treatment model (2), with the intercept α_0 set to -0.9. Thus, approximately 30% of the simulated subjects received the treatment. The outcome variable Y was generated using the linear model outlined in Eq. (1), incorporating an error term $\sim N(0, 1)$. β_1 was set at 0.5 in "Imbalance and bias" section and 1 in "Model dependence" section.

(iii) In each simulated data set, we computed the PS for every subject using a logistic regression model. We then applied a nearest-neighborhood matching algorithm to pair each treated subject with a control subject based on the logit of the PS. This matching was performed without replacement, utilizing a caliper width equal to c times the standard deviation of the logit PS, where c was selected from the set of values $\{20, 1, 0.2, 0.02, 0.002, 0.0002\}$. Matching algorithm was implemented in `R/MatchIt`.

Imbalance and bias

At each matching caliper size, we calculated various metrics for assessing imbalance, including the number of the matched pairs, SMD of confounder X_3, and the multivariate Mahalanobis distance between group means of all confounders in the treated and untreated groups. This analysis was conducted on 5000 randomly generated samples, with the same calculations repeated in each sample. Subsequently, we averaged these measurements across the 5000 replicates. Additionally, we computed the proportion of SMDs of X_3 that exceeded 0.1 in the absolute values. We anticipated that the averaged SMD of X_3 would effectively capture any chance imbalance and thus converge toward 0 after the initial balancing was achieved. In contrast, we anticipated that the average of Mahalanobis metrics would initially decrease until a caliper size of 0.2 × the standard deviation of the logit PS was reached, after which they would increase. The proportion of absolute SMDs of X_3 larger than 0.1 was also expected to follow this trend. These increasing trends reflect the increasing likelihood of observing substantial chance imbalances due to a continued decrease in sample size after the initial balancing was established.

In addition to calculating these imbalance metrics at each caliper size, we performed regression analyses. Three different models were fitted: an unadjusted model with the treatment indicator A (referred to as "$\mathcal{M}(A)$"), a multiple regression model including A, linear terms of X_4, and X_5 (referred to as the "$\mathcal{M}(A, X_4, X_5)$"), a model including A, linear terms of X_1 through X_5 (referred to as the "$\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$"), and finally, we extracted estimates of the treatment effect from each model, along with model-based standard errors and robust sandwich standard errors. These effect estimates and variance estimates were then averaged across 5000 replicates. Additionally, we computed the empirical variance estimate for each effect estimator by calculating the variance of 5000 effect estimates. We anticipate that $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ would outperform both $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$, displaying the smallest bias and variance estimates. When considering the commonly recommended caliper size of 0.2 × the standard deviation of the logit PS, we expected that even though both $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$ were misspecified, they will still produce nearly unbiased effect estimations.

Model dependence

We used a new set of coefficients and a more complex outcome model that incorporates quadratic and interaction terms to generate the outcome Y (refer to the Web Appendix 3 for details). Within the framework of this complex outcome model, we evaluated the performance of misspecified models for effect estimation. Two different models were employed: $\mathcal{M}(A)$ and $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$. We then extracted estimates of the treatment effect from each model, along with model-based standard errors. These effect estimates and variance estimates were subsequently averaged across 5000 replicates. Additionally, we computed the empirical variance estimate for each effect estimator by calculating the variances of the 5000 effect estimates. We anticipate that with a caliper size of 0.2 times the standard deviation of the logit PS, both $\mathcal{M}(A)$ and $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ would yield nearly unbiased effect estimates. To examine the claim that even inexact matching can still make parametric analyses less sensitive to model misspecification, we also fitted $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ in unmatched data and computed the averaged estimate of β_1.

Re-assessing the type of randomized design PSM emulates in previous study

We follow the same data generation approach outlined in section (5.2) in the previous study [10]. We first generate 5000 treated and 5000 untreated subjects (i.e., $A \sim Binom(0.5)$). For each of two covariates X_1 and X_2, we randomly and independently draw control units from $Uniform(1, 6)$ and treated units from $Uniform(0, 5)$. We generate the outcome as $Y_i = 2A + X_1 + X_2 + \epsilon_1$, where $\epsilon_i \sim N(0, 1)$. Thus, the effects of X_1 and X_2 on the treatment are expected to be the same and proportional to their effects on the outcome. i.e., $\sin(\theta) = 0$. We repeat the simulation 5000 times and in each simulation, we performed the same analyses as outlined in Imbalance and bias section. For each matched data, we used generalized estimation equation method to compute the intraclass correlation coefficients within the matched pairs. The simulation code is provided in supplemental material (Web Appendix 2).

Simulation results

Figure 1 illustrates the simulation results for imbalance and biases in the scenario involving low within-pair correlation, where the *sine* distance exceeds 0.8 and PSM approximates a CRD. Firstly, as the caliper size diminishes, the Mahalanobis distance initially decreases until it reaches the optimal caliper size (0.2× the standard deviation of logit PS), and then increases as the caliper size continues to shrink, and the sample size decreases (see Fig. 1A). The proportion of absolute SMD greater than 0.1 for X_3 follows a similar pattern (Fig. 1B). SMD of X_3 decreases until reaching the optimal caliper size and then stabilizes near zero thereafter (Fig. 1C). This confirms that metrics like the Mahalanobis distance only reflects the increasing variability in chance imbalance and a higher likelihood of observing larger chance imbalance as the sample size decreases. Only metrics that retain the direction of differences, such as SMD, can accurately measure this chance imbalance. Secondly, we can also observe that $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$ produce biased estimates until confounders are balanced at the optimal caliper size, and PSM approximates a randomized design (Fig. 1D). Given that the true outcome model includes linear terms for five confounders, $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ consistently produces unbiased results at all caliper sizes. When either the regression model is correctly specified or matching is performed correctly (even with incorrect models), we can consistently obtain unbiased results (Fig. 1D). Lastly, it's worth noting that the model-based standard errors for both $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$ do not align precisely with their empirical standard errors,

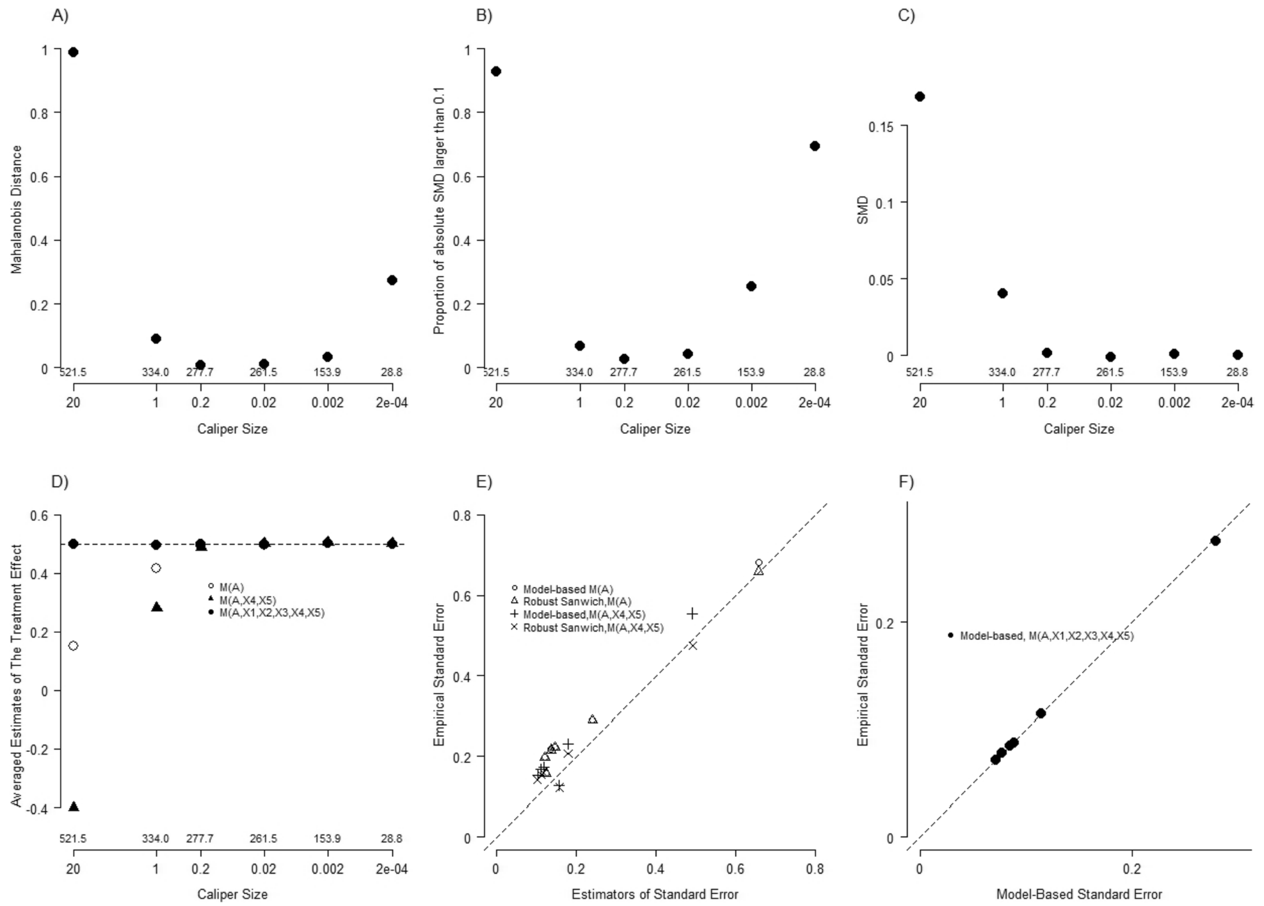

Fig. 1 Imbalance and Bias for the *sine* distance > 0.8. **A** The trend of the mahalanobis distance with shrinking caliper size; (**B**) The trend of the proportion of absolute SMD of X_3 larger than 0.1; (**C**) The trend of SMD of X_3; (**D**) The trend of estimators of PATT; (**E**) The concordance between empirical standard error and model-based/Robust Sandwich estimators for $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$; (**F**) The concordance between empirical standard error and model-based estimators for $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$

and robust sandwich estimators do not show significant improvement (Fig. 1E). The model-based standard errors from $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ align well with its empirical standard errors (Fig. 1F).

Figure 2 illustrates the simulation results for imbalances and biases in a scenario where the within-pair correlation is high (i.e., the *sine* distance is less than 0.2 and PSM approximates a randomized blcok design). We observe a consistent pattern of imbalance and bias across different aspects (Fig. 2A-D). Both $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$ exhibit bias until the caliper size reaches the optimal level. In contrast, the correctly specified $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ remains unbiased throughout the process. Notably, robust sandwich standard error estimates for $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$ perform better than model-based standard error estimates after the caliper size reaches the optimal level (Fig. 2E and Table 2 in Web Appendix 4). The model-based standard errors of $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ closely align with their empirical standard errors (Fig. 2F).

Figure 3 illustrates the sensitivity of PSM to model misspecification in our simulation results. We observe a consistent pattern in imbalance metrics (Fig. 3A-C). Both misspecified models show bias until the matching caliper size reaches its optimal level (Fig. 3D). Notably, model-based standard errors for $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ closely match empirical standard errors across all caliper sizes. However, this alignment is poor for $\mathcal{M}(A)$, although the robust sandwich variance estimator improves the alignment at and after the optimal caliper size (Fig. 3 E and F and Table 3 in Web Appendix 4). Moreover, the averaged estimate of β_1 obtained using $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$ in the unmatched dataset is 1.37, indicating a considerably higher bias compared to when the same model is applied in a poorly matched PSM design, employing a large caliper size of 20 × the standard deviation of the logit PS (Fig. 3D). This reaffirms the previous conclusion [26] that matching, even an imperfect one, can effectively reduce the sensitivity of parametric

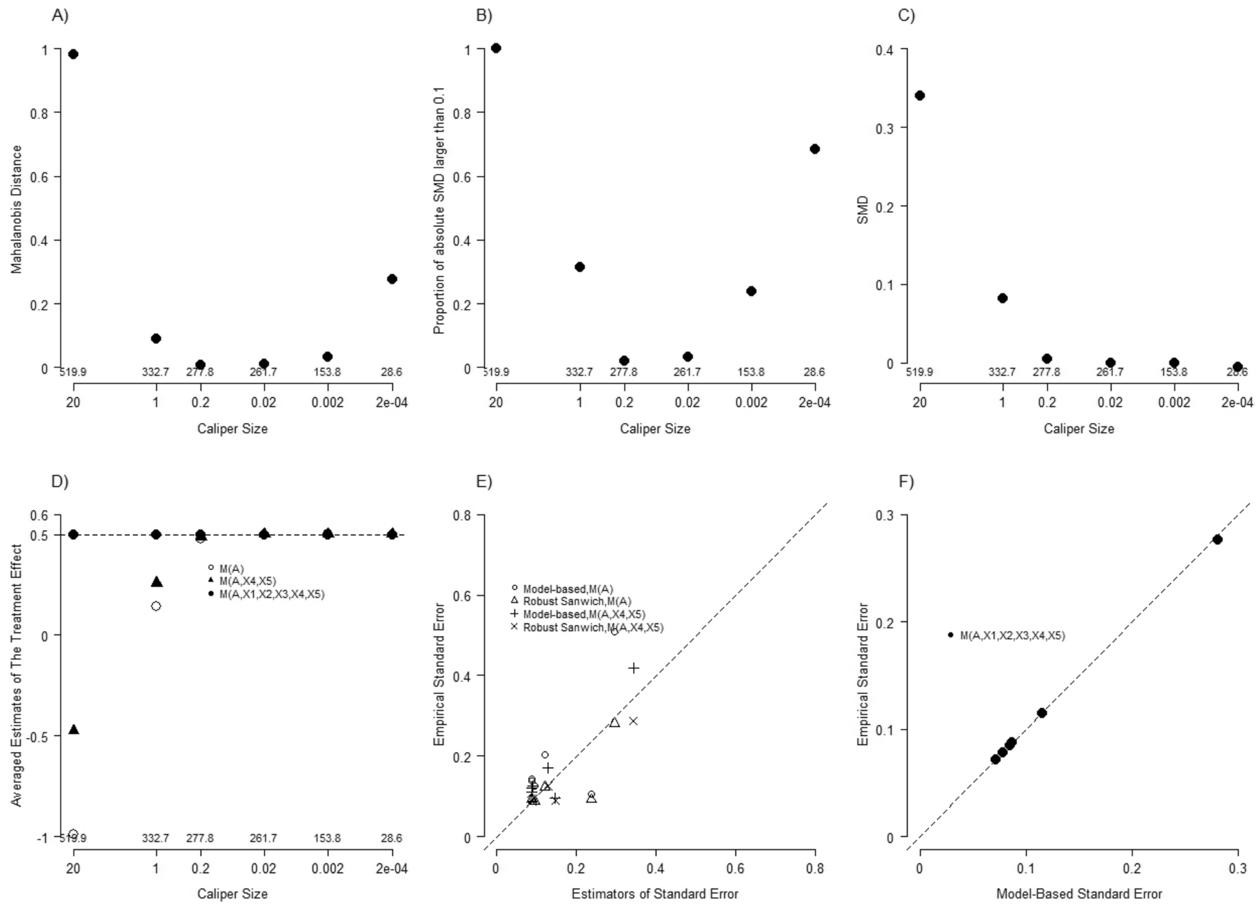

Fig. 2 Imbalance and Bias for the *sine* distance < 0.2. **A** The trend of the mahalanobis distance with shrinking caliper size; (**B**) The trend of the proportion of absolute SMD of X_3 larger than 0.1; (**C**) The trend of SMD of X_3; (**D**) The trend of estimators of PATT; (**E**) The concordance between empirical standard error and model-based/Robust Sandwich estimators for $\mathcal{M}(A)$ and $\mathcal{M}(A, X_4, X_5)$; (**F**) The concordance between empirical standard error and model-based estimators for $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$

modeling to model misspecification. We further examined imbalance, bias, and model dependence using 3 correlated normally distributed X_1, X_2, X_3 ($\rho = 0.5$) and 2 correlated binary X_4, X_5 ($\rho = 0.3$) and the results are consistent (details in Web Appendix 5).

Figure 4 illustrates the simulation results for the setting used in the previous study [10]. A consistent pattern of imbalance and bias is observed across various metrics (Fig. 4A-D). Firstly, as the caliper size decreases, the Mahalanobis distance initially declines until reaching the optimal caliper size (0.2× the standard deviation of the logit PS). Beyond this point, it increases due to reduced sample size (Fig. 4A). The proportion of absolute SMDs exceeding 0.1 for X_1 follows a similar trajectory (Fig. 4B). The SMD for X_1 decreases up to the optimal caliper size and then stabilizes near zero (Fig. 4C). This indicates that metrics like the Mahalanobis distance reflect the growing variability in chance imbalance and the heightened likelihood of observing larger chance imbalances as sample size decreases. The unadjusted effect estimator $\mathcal{M}(A)$ exhibits bias until the caliper size reaches its optimal value (Fig. 4E). Furthermore, Fig. 4D demonstrates a strong within-pair intraclass correlation coefficient ($\rho_2 > 0.7$) for PSM. Thus, under this setting, PSM mimics a RBD rather than a CRD as claimed. The estimated coefficients $\boldsymbol{\alpha_1} \approxeq \langle 0.47, 0.47 \rangle$ are proportional to $\boldsymbol{\beta_1} = \langle 1, 1 \rangle$, yielding a *sine* distance of 0. As discussed in The type of randomized design PSM emulates section, the proportionality of these coefficient sets and the zero *sine* distance are linked to strong intraclass correlation.

In summary, in line with the balancing score property of PS, the observed imbalance in PSM primarily arises due to chance, which cannot be adequately captured by absolute distance metrics like the Mahalanobis distance. This chance-driven imbalance eventually averages to zero when the matching caliper size is optimal, thereby not biasing effect estimation when using misspecified models. PSM reduces sensitivity to model misspecification in

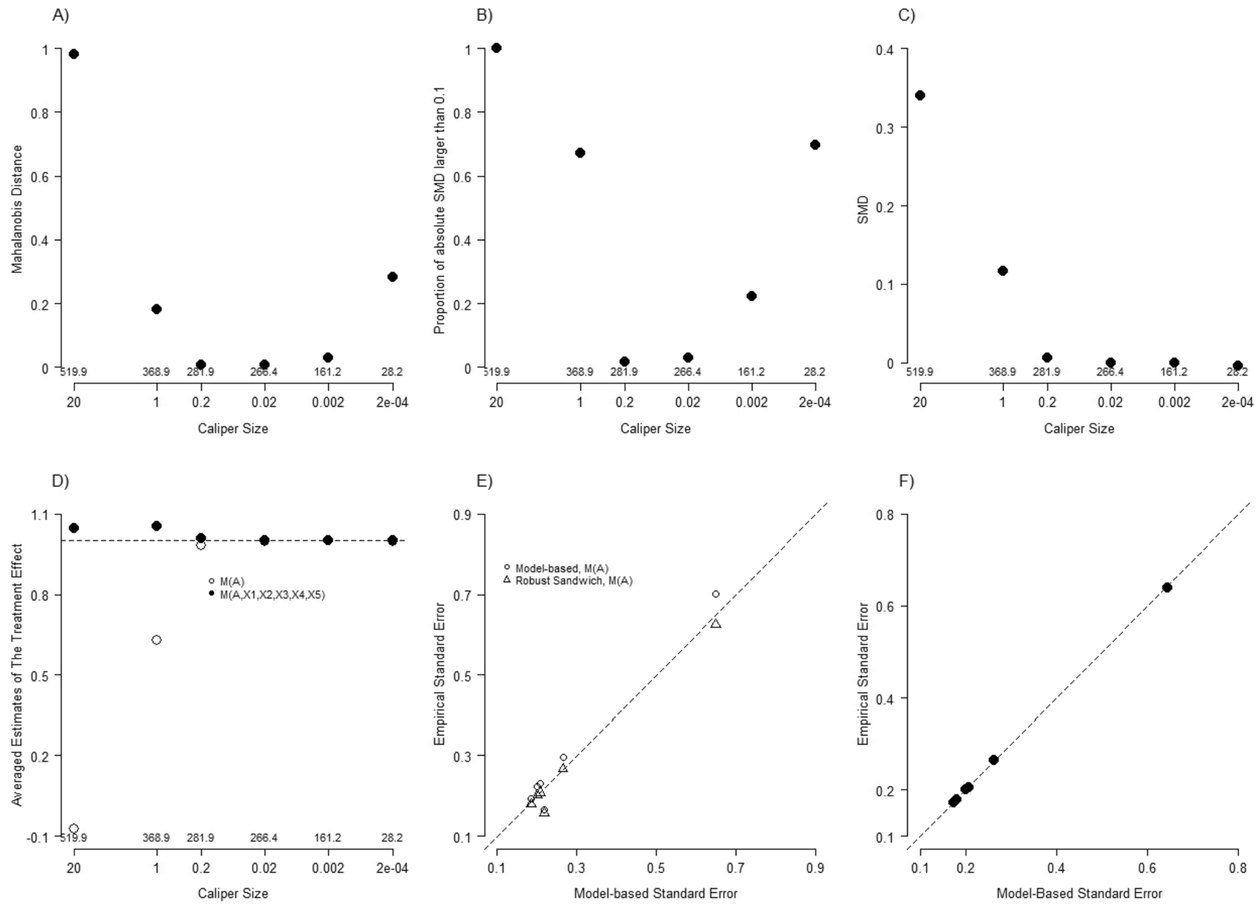

Fig. 3 Sensitivity to Model Dependence. **A** The trend of the mahalanobis distance with shrinking caliper size; (**B**) The trend of the proportion of absolute SMD of X_3 larger than 0.1; (**C**) The trend of SMD of X_3; (**D**) The trend of estimators of PATE; (**E**) The concordance between empirical standard error and model-based estimators for $\mathcal{M}(A)$; (**F**) The concordance between empirical standard error and model-based estimators for $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$

post-matching analysis. Under PSM, model-based standard errors for misspecified models generally fail to align with empirical standard errors, except for those from $\mathcal{M}(A, X_1, X_2, X_3, X_4, X_5)$. It's essential to note that selectively choosing the best result from a multitude of regression models is not only biased but also unnecessary.

Discussion

In this study, we have successfully validated the theoretical properties of PS in the context of matching designs. Specifically, when PSM approaches exact matching, it effectively balances confounding variables between comparison groups, and any observed imbalances are merely due to chance. Additionally, our findings confirm that PSM design mitigates the sensitivity to model misspecification in post-matching analysis, a characteristic described by Guo and Rothenhäusler for a matching design [26]. Both a simple group mean difference and regression adjustment using linear terms of matching factors can accurately estimate PATT. Our findings stand in contrast to the previously identified PSM paradox [10]. This paradox suggests that model dependence and statistical bias should increase as units are pruned. We conclude that this discrepancy primarily arises from the use of inappropriate metrics for assessing imbalance and bias in the previous study [10]. Consequently, there is no valid concern that should deter us from employing PSM in comparative effectiveness research.

Differing from other highly cited studies that emphasize population treatment effect [4], King and Nielsen [10] have stated that sample treatment effect is the causal interest. SATE or SATT represents the treatment effect exclusively within the context of the available sample data, rather than being applicable to the broader target population. Upon closer examination of their simulation design, it becomes evident that King and Nielsen are, in fact, targeting a population treatment effect. In their simulation study, the causal parameter is represented by

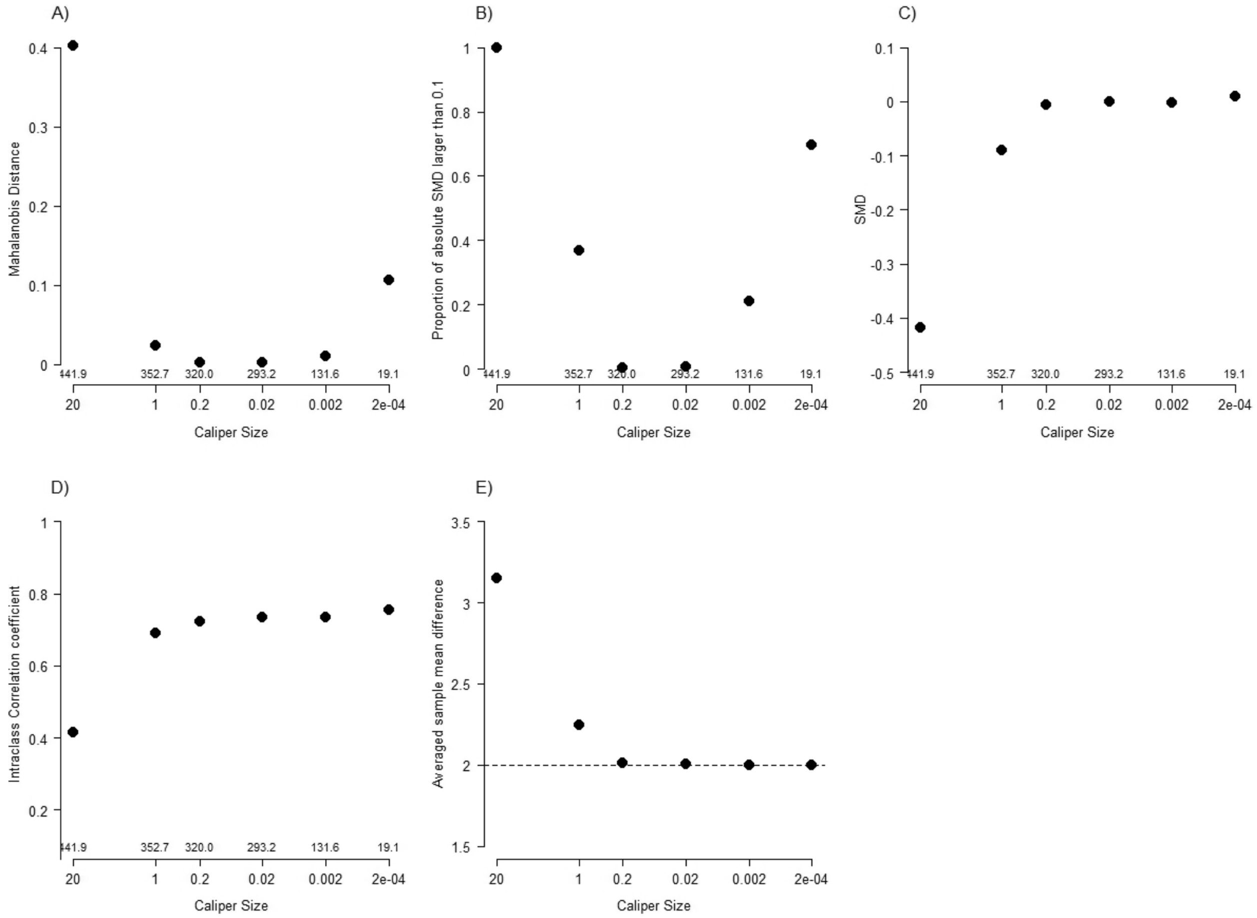

Fig. 4 Assessing PSM using the previous simulation setting [10]. **A** The trend of the mahalanobis distance with shrinking caliper size; (**B**) The trend of the proportion of absolute SMD of X_1 larger than 0.1; (**C**) The trend of SMD of X_1; (**D**) The trend of Intraclass Correlation Coefficient; (**E**) The trend of unadjusted effect estimator of PATT

the coefficient of the treatment indicator in an additive model that does not include treatment-by-confounder interactions. This is a conventional approach for representing a homogeneous population treatment effect, and they draw different random samples to calculate the average effect estimate. This approach aligns with the typical methodology for computing the expected value of an effect estimator for a population treatment effect, similar to what we have done in this study and other studies [4]. Therefore, it is important to note that the previous study [10] has not formally demonstrated the bias in estimating sample treatment effects in a PSM design. Even if such a demonstration was made, it would not be sufficient grounds to dismiss the utility of PSM because PSM was initially proposed to target population treatment effects and has been substantiated through simulation studies in this context [3, 4].

We have also demonstrated that the prior findings concerning the eventual increase in imbalance as units are progressively pruned [10, 20], a significant consequence of the PSM paradox, can be attributed to the inappropriate imbalance metrics that were employed. As extensively discussed by Rosenbaum [16], the covariate values of treated and untreated subjects matched on the same PS often exhibit different covariate values, which occur by chance and can swing in either direction. Consequently, when the number of matched pairs is substantial, these differences tend to average towards zero in PSM design. When matched samples are finite, the imbalances, when averaged across all matched samples, converge toward zero, or in other words, the expected value of imbalance is zero. Thus, using the average pairwise Mahalanobis absolute distance in covariate space between comparison groups as a measure of imbalance in a PSM design misinterprets the balancing property of the PS. Note that the Mahalanobis absolute distance remains a useful tool for assessing imbalance in other CM designs when matching is inexact.

Chance imbalances do not predict treatment status and should not introduce bias when estimating population treatment effects [16]. This topic is widely discussed in the context of randomized designs, with differing philosophical perspectives on the role of chance imbalance [30]. One viewpoint holds that balance in a randomized study is essential for valid inference. However, Senn [31] argued that randomization does not guarantee balance in an individual CRD study. Even with matching or blocking, unmatched or unblocked variables may remain imbalanced between arms. Nevertheless, chance imbalance does not invalidate point estimates (e.g., sample mean differences) or conventional hypothesis testing (e.g., two-sample t-tests). Any deviation between the point estimate and the true value reflects uncertainty or variability, not statistical bias. We can always adjust for unbalanced variables predicative of outcome in an analysis of covariance model (ANCOVA) to increase efficiency.

King and Nielsen [10] showed increased model-dependence and statistical bias connected with the PSM paradox using a cherry-picking estimation procedure. However, we argue that evaluating model dependence of PSM based on biased cherry-picking procedure is not appropriate. As shown in our simulation and other studies [21, 26], PSM reduces the sensitivity of model misspecification because we can fit a well-considered model for estimating the treatment effect with satisfactory precision and efficiency, even if the chosen model could be misspecified. The key advantage of PSM, as a matching design, is that it frees us from the need to explore all possible models and search for the best results in post-matching analysis, which is a biased practice that should be avoided. This practice can lead to the most dramatic effect estimate even in randomized studies [32]. The reduction on model dependence by PSM can be explained by the fact that matching can balance any function of $\mathbf{X}$, making A and any function of $\mathbf{X}$ approximately orthogonal in the matched sample. Thus, the inclusion or exclusion of a nearly orthogonal predictor has negligible effects on the other regression coefficients based on least-square theory [26]. This resembles similar practice in analyzing a CRD. We can perform unadjusted analysis and also adjusted regression with gains in efficiency. However, what variables should enter into model and what forms they should take must be determined at the design stage.

Conclusions

This study focuses on the question, "Is PSM a valid design?" rather than "Is PSM superior to other CM designs?". Debates about which design, PSM or CM, is superior in specific settings are always legitimate, considering the trade-off between statistical bias and efficiency. PSM, like any statistical method, is not a universal solution. It has notable limitations, including:

- PSM offers a practical approach to addressing the curse of dimensionality in matching designs. However, it may fail under substantial unmeasured confounding.
- Achieving unbiased treatment effect estimation in PSM requires correctly specifying the PS model [19]. Although matching designs are robust to misspecification of the outcome model, incorrect PS estimation undermines its balancing score and ignorability properties. Flexible machine learning methods, such as generalized boosted regression, can help by accounting for nonlinearities and covariate interactions.
- PSM designs encompass both matching (design stage) and effect estimation or hypothesis testing (analysis stage). Conventional model-based inference for analyzing randomized designs, often adopted based on the premise that PSM mimics randomized designs, may be invalid. Further research is needed on variance estimators to address this issue [22].
- Discussions of the PSM paradox typically involve continuous outcomes and mean differences. However, clinical studies often focus on binary or time-to-event outcomes, which target non-collapsible effect measures such as odds ratios and hazard ratios. In these cases, marginal effects (averaged over the population) and conditional effects (conditioned on population characteristics) may differ. Additional considerations of causal quantities are necessary for analyzing PSM designs [28, 33].

Nevertheless, it's important to recognize that the PSM paradox should not overly concern researchers.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02481-w.

Supplementary Material 1

Acknowledgements
Not applicable.

Authors' contributions

FW proposed the study concept and simulation design and drafted the manuscript.

Funding
Not applicable.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 16 September 2024 Accepted: 20 January 2025
Published online: 29 January 2025

References

1. Pearl J. The Foundations of Causal Inference. Sociol Methodol. 2010;40(1):75–149.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
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(1):171–84.
4. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014;33(6):1057–69.
5. Dehejia RH, Wahba S. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. J Am Stat Assoc. 1999;94(448):1053–62.
6. Dehejia RH, Wahba S. Propensity Score-Matching Methods for Nonexperimental Causal Studies. Rev Econ Stat. 2002;84(1):151–61.
7. Zhao Z. Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. Rev Econ Stat. 2004;86(1):91–107.
8. Wan F. Matched or unmatched analyses with propensity-score-matched data? Stat Med. 2019;38(2):289–300.
9. Abadie A, Imbens GW. Matching on the Estimated Propensity Score. Econometrica. 2016;84(2):781–807.
10. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Political Anal. 2019;27(4):435–54.
11. Wyler Von Ballmoos MC, Almassi GH. Commentary: Delayed gratification and optimism bias: Navigating quality and quantity of life with revascularization in patients with ischemic cardiomyopathy. J Thorac Cardiovasc Surg. 2021;161(3):1032–3.
12. Sceats LA, Trickey AW, Morris AM, Kin C, Staudenmayer KL. Nonoperative Management of Uncomplicated Appendicitis Among Privately Insured Patients. JAMA Surg. 2019;154(2):141.
13. Sareen V, Ho D. Preoperative Medical Consultation-Questioning a Long-Standing Practice. JAMA Intern Med. 2023;183(9):1033–4.
14. 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(2):150–61.
15. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037–49.
16. Rosenbaum PR. Design of observational studies. 2nd ed. Springer series in statistics. Cham: Springer; 2020.
17. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
18. Wan F. An interpretation of the properties of the propensity score in the regression framework. Commun Stat Theory Methods. 2021;50(9):2096–105.
19. Lenis D, Ackerman B, Stuart EA. Measuring model misspecification: Application to propensity score methods with complex survey data. Comput Stat Data Anal. 2018;128:48–57.
20. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. Am J Epidemiol. 2018;187(9):1951–61.
21. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Anal. 2007;15(3):199–236.
22. Austin PC, Cafri G. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. Stat Med. 2020;39(11):1623–40.
23. Wan F, Sutcliffe S, Zhang J, Small D. Does matching introduce confounding or selection bias into the matched case-control design? Observational Stud. 2024;10(1):1–9.
24. Wan F. Conditional or unconditional logistic regression for frequency matched case-control design? Stat Med. 2022;41(6):1023–41.
25. Wan F, Colditz GA, Sutcliffe S. Matched Versus Unmatched Analysis of Matched Case-Control Studies. Am J Epidemiol. 2021;190(9):1859–66.
26. Guo K, Rothenhäusler D. On the statistical role of inexact matching in observational studies. Biometrika. 2023;110:631–44.
27. Rubin DB. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics. 1973;29(1):185.
28. Wan F, Mitra N. An evaluation of bias in propensity score-adjusted nonlinear regression models. Stat Methods Med Res. 2018;27(3):846–62.
29. Wan F. A Cautionary Note on Using Propensity Score Calibration to Control for Unmeasured Confounding Bias When the Surrogacy Assumption Is Absent. Am J Epidemiol. 2024;193(2):360–9.
30. Rosenberger WF, Sverdlov O. Handling covariates in the design of clinical trials. Stat Sci. 2008;23(3):404–19.
31. Senn S. Seven myths of randomisation in clinical trials. Stat Med. 2013;32(9):1439–50.
32. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Stat Med. 2008;27(23):4658–77.
33. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078–94.

Publisher's Note