Testing Biased Randomization Assumptions and Quantifying Imperfect Matching and Residual Confounding in Matched Observational Studies

Kan Chen, Siyu Heng, Qi Long, and Bo Zhang

ABSTRACT
One central goal of design of observational studies is to embed nonexperimental data into an approximate randomized controlled trial using statistical matching. Despite empirical researchers’ best intention and effort to create high-quality matched samples, residual imbalance due to observed covariates not being well matched often persists. Although statistical tests have been developed to test the randomization assumption and its implications, few provide a means to quantify the level of residual confounding due to observed covariates not being well matched in matched samples. In this article, we develop two generic classes of exact statistical tests for a biased randomization assumption. One important by-product of our testing framework is a quantity called residual sensitivity value (RSV), which provides a means to quantify the level of residual confounding due to imperfect matching of observed covariates in a matched sample. We advocate taking into account RSV in the downstream primary analysis. The proposed methodology is illustrated by re-examining a famous observational study concerning the effect of right heart catheterization (RHC) in the initial care of critically ill patients. Code implementing the method can be found in the supplementary materials.

1. Introduction
1.1. Statistical Matching and Randomization-Based Outcome Analysis
In an observational study of a treatment’s effect on the outcome, units involved in the study may differ systematically in their observed pretreatment covariates, thus invalidating a naive comparison between the treated and control groups. Statistical matching and subclassification are commonly used nonparametric tools to adjust for covariates. For a binary treatment, the ultimate goal of statistical matching is to embed observational data into an approximate randomized experiment by designing a treated group and a matched control (or comparison) group that are comparable in their observed covariates (Rosenbaum 2002, 2010; Stuart 2010; Bind and Rubin 2019).

One widely used downstream outcome analysis method for matched data is randomization-based inference as if the matched data comes from a randomized controlled trial (RCT). In our view, this practice is appealing for two reasons. First, randomization inference (possibly with regression adjustment) is commonly used in analyzing RCT data; hence, using a randomization-based procedure to analyze matched data fits into the big picture of statistical matching, that is, to embed observational data into an experiment. Second, the sensitivity analysis framework complementing the primary randomization-based outcome analysis is well understood and developed. The Rosenbaum-bounds-type sensitivity analysis (Rosenbaum 2002, 2010; DiPrete and Gangl 2004) allows the treatment assignment probability in each matched pair (or set) to deviate from the randomization probability up to a degree controlled by a parameter $\Gamma$ and outputs a bounding $p$-value under such deviation. Furthermore, randomization-based inferential methods can be applied to testing both Fisher’s sharp null hypothesis (Rosenbaum 2002, 2010) and Neyman’s weak null hypothesis (Wu and Ding 2020). Sensitivity analysis methods have also been developed for both sharp and weak null hypotheses (Rosenbaum 2002, 2010; Fogarty 2020).

1.2. The Randomization Assumption
In a typical randomization-based downstream outcome analysis of matched-pair data, researchers make the following Randomization Assumption (Rosenbaum 2002, 2010):

**Assumption (Randomization Assumption, Stated Informally).** The treatment assignments across all matched pairs are assumed to be independent of each other and that the treatment is randomly assigned in each matched pair, that is, probability of the first unit receiving treatment and the other control is the same as the second unit receiving treatment and the first control.
The randomization assumption (RA) entails two aspects: (i) independence among matched pairs and (ii) random treatment assignment in each matched pair. It is the single most important assumption as it enables researchers to treat observational data after statistical matching as if the data were from a randomized controlled experiment (Dehejia and Wahba 2002; Rosenbaum 2002, 2010). Researchers make analogous assumptions when analyzing observational data after 1-to-k matching, full matching (Heng et al. 2021; Zhang, Mackay, and Baiocchi 2022), and matched-pair clustered designs (Hansen, Rosenbaum, and Small 2014; Zhang et al. 2021a).

The RA clearly holds by design in an RCT. It also holds if matched pairs are independent of each other and two units in the same pair have the same propensity score. In practice, the true propensity score for each unit is at best known up to estimation uncertainty if one happens to correctly specify the propensity score model and at worst unfathomable. Moreover, the matching algorithm may create weak dependence among matched pairs or sets so that the randomization assumption is at best an approximation of the reality.

Researchers perform randomization-based outcome analysis for assorted matched data, not restricted to propensity-score-matched (PSM) data. In fact, as shown by Rosenbaum and Rubin (1985), propensity score matching often underperforms more sophisticated matching algorithms that combine metric-based and propensity-score-based matching in balancing observed covariates. Indeed, many modern statistical matching algorithms have moved beyond PSM; some notable examples include network-flow-based optimal matching and its many variants (Rosenbaum 1989, 2002, 2010; Zhang et al. 2021b, among others), coarsened exact matching (Iacus, King, and Porro 2011), mixed-integer-programming-based algorithms (Zubizarreta 2012), and genetic matching algorithm (Diamond and Sekhon 2013), among others.

### 1.3. Justifications for Randomization Inference: Informal and Formal Diagnostics

To justify using a randomization-based inferential procedure, researchers often perform informal diagnostics based on metrics like the standardized mean differences (SMDs) to examine the covariate balance after matching (Silber et al. 2001, Franklin et al. 2014, Austin and Stuart 2015) or formal statistical procedures to test the equality of covariate distributions in the treated and matched control groups (e.g., Rosenbaum’s (Rosenbaum 2005) crossmatch test). However, there is a gap between equality of covariate distributions and the RA: the downstream randomization inference relies solely on the RA and need not assume sampling covariates or potential outcomes from some superpopulation. A more detailed review and related discussion can be found in Appendix A.

More recently, Gagnon-Bartsch and Shem-Tov (2019) developed the Classification Permutation Test (CPT) that can be adapted to testing the RA as follows. First, train a classifier \( \hat{f} \) using any classification tools (e.g., logistic regression, support vector machine, random forests, or an ensemble of them) with treatment status \( Z \) as the label and observed covariates \( X \) as predictors, and predict treatment indicators for all matched-pair data using \( \hat{f} \). Denote the predicted treatment indicators by \( \hat{Z}_{11}, \ldots, \hat{Z}_{12} \) and define the test statistic \( T = \sum_{i=1}^{r} \sum_{j=1}^{2} Z_{ij} \hat{Z}_{ij} \). Next, for \( m = 1, \ldots, MC \), randomly permute two treatment indicators within each pair, denoted as \( Z^{(m)}_{11}, \ldots, Z^{(m)}_{12} \), retrain the classifier \( \hat{f}^{(m)} \) using the same covariates data \( X \) but permuted treatment indicators \( Z^{(m)}_{11}, \ldots, Z^{(m)}_{12} \), and then re-predict treatment assignments using \( \hat{f}^{(m)} \). Denote by \( \hat{Z}_{11}^{(m)}, \hat{Z}_{12}^{(m)} \) the prediction at iteration \( m \). To test the randomization assumption, it then suffices to compare the test statistic \( T \) to the null distribution generated by \( T^{(m)} = \sum_{i=1}^{r} \sum_{j=1}^{2} Z_{ij}^{(m)} \hat{Z}_{ij}^{(m)} \), \( m = 1, \ldots, MC \). In this way, the CPT procedure yields an exact \( p \)-value for testing the randomization assumption. Similar Fisherian-style permutation strategies are also used in Branson (2020) and Branson and Keele (2020) to deliver an exact test for the RA.

### 1.4. Moving Beyond the Randomization Assumption: Testing a Biased Randomization Assumption and Quantifying Residual Confounding

A powerful test of the RA is desirable as it closely examines a most important premise for downstream statistical inference; however, a test of the RA by itself may not be comprehensive enough to capture the full picture of the study design. Take as an example Gagnon-Bartsch and Shem-Tov’s (2019) reanalysis of Heller, Rosenbaum, and Small’s (2010) matched-pair data. (Gagnon-Bartsch and Shem-Tov 2019, Section 4.4) rejected the RA at the 0.05 level using the CPT procedure and concluded that “the covariates can predict the treatment assignment better than under random assignment.” As one carefully examines the covariate balance of Heller, Rosenbaum, and Small’s (2010) matched-pair data, however, no two-sample \( t \)-test, Wilcoxon signed rank test, or Kolmogorov-Smirnov test is statistically significant at the 0.1 level for any of the observed covariates in the treated and matched control groups, suggesting that there is likely to be little residual confounding due to imperfect matching on observed covariates. It is also unclear to what extent a minor deviation from the RA would affect the downstream outcome analysis. From a practical perspective, with a dataset as well-matched as that in Heller, Rosenbaum, and Small (2010), few empirical researchers would redo the sometimes computationally intensive statistical matching. A powerful test of the RA may thus have an unintended consequence of discouraging empirical researchers from adopting and reporting such formal diagnostics. In our opinion, what currently are lacking in the literature are 3-fold: (i) a measure to quantify the level of residual imbalance in observed covariates, an arguably more important practical matter than testing and rejecting the RA by itself, (ii) a means to systematically incorporate “recalcitrant” residual covariate imbalance, if there is any, into the outcome analysis, and (iii) simulation results that relate the quality of statistical matching to the performance of outcome analysis.

### 1.5. Our Contribution

This article aims to start filling in these gaps. We propose two generic classes of exact statistical tests for a relaxed ver-
2. Randomization and Biased Randomization Assumption

We consider the setting of a typical matched cohort study. Suppose that the study has access to $N_t$ treated units and a large reservoir of $N_c$ control units so that there are $N_t + N_c = N$ units in total. Without loss of generality, we assume $N_c > N_t$ (if not, switch the role of treated and control units) and $N_t$ is often much larger than $N_c$. Each of the $N$ units is associated with a treatment indicator $Z_n$ and a $q$-dimensional vector of observed covariates $x_n$, $n = 1, \ldots, N$. Researchers match some subset of the treated units with control units using any statistical matching procedure $M$ and produce $I \leq N_t$ matched pairs of two units, one treated and the other control. We use $ij$, $i = 1, \ldots, I$, $j = 1, 2$, to index the $j$-th unit in the $i$-th matched pair. Let $x_{ij}$ denote unit $ij$’s observed covariates and $Z_{ij}$ its treatment status so that the unit with $Z_{ij} = 1$ is the treated unit and with $Z_{ij} = 0$ the control unit. Finally, we collect the observed covariates information of $2I$ matched units in the matrix $X$. We assume no unmeasured confounders.

The randomization assumption in a downstream, finite-sample, randomization-based outcome analysis can be formally stated as follows:

**Assumption 1 (Randomization Assumption in Matched-Pair Studies, Stated Formally).** Treatment assignments across matched pairs are assumed to be independent of each other, with

$$P(Z_{i1} = 1, Z_{i2} = 0|X,M) = P(Z_{i1} = 0, Z_{i2} = 1|X,M) = 1/2,$$

for $i = 1, \ldots, I$.  

Statistical matching can largely remove overt bias from observed covariates (Rosenbaum 2002, 2010); however, some residual confounding from observed covariates may persist in the matched sample due to imperfect matching, and this motivates a biased randomization assumption:

**Assumption 2 (Biased Randomization Assumption).** Treatment assignments across matched pairs are assumed to be independent of each other, with

$$\Gamma^{-1} \leq \frac{P(Z_{i1} = 1, Z_{i2} = 0|X,M)}{P(Z_{i1} = 0, Z_{i2} = 1|X,M)} \leq \Gamma,$$

for $i = 1, \ldots, I$, and some $\Gamma \in [1, \infty)$. 

Assumption 2 is the basis for the Rosenbaum-bounds sensitivity analysis framework (Rosenbaum 2002, 2010). It uses one parameter $\Gamma$ to control the maximum degree to which the residual confounding biases the treatment assignment probability in each matched pair. We stress again that we assume no unmeasured confounding in this article and are only interested in assessing residual confounding from observed covariates. In the presence of unmeasured confounding, $\Gamma$ can be arbitrarily large, and Assumption 2 becomes untestable.

3. Sample-Splitting Classification Permutation Test (SS-CPT)

Our first proposal to test Assumption 2 is a simple modification of the CPT. First, randomly split $I$ matched pairs into two nonoverlapping index sets $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$, and train a classifier $\hat{f}_1$ using observed covariates $X^{(1)} = \{X_{ij}, i \in \mathcal{T}^{(1)}, j = 1, 2\}$ as predictors and treatment status $Z^{(1)} = \{Z_{ij}, i \in \mathcal{T}^{(1)}, j = 1, 2\}$ as labels. Next, apply the classifier $\hat{f}_1$ to the observed covariates $X^{(2)} = \{X_{ij}, i \in \mathcal{T}^{(2)}, j = 1, 2\}$ and let $\hat{T}_{1-2} = \{f(x_{ij}), i \in \mathcal{T}^{(2)}, j = 1, 2\} \subseteq [0,1]^{2\mathcal{T}^{(2)}}$ denote the predicted scores. Lastly, let $g_{ij} : [0,1]^{2\mathcal{T}^{(2)}} \to \mathbb{R}$ denote a generic function and form the test statistic $T_{1-2} = \sum_{i<j \in \mathcal{T}^{(2)}} Z_{ij}g_{ij}(\hat{f}_{1-2})$, which can then be used to test the biased RA restricted to matched pairs $i \in \mathcal{T}^{(2)}$ for a fixed $\Gamma$ value according to Proposition 1 below.

**Proposition 1 (Bounding $p$-value).** Let $T$ be a test statistic of the form $T = \sum_{i=1}^I \sum_{j=1}^2 Z_{ij}q_{ij}$, where $q_{ij}$ is some fixed score based on the observed covariates $X = \{x_{ij} : i = 1, \ldots, I, j = 1, 2\}$. For $i = 1, \ldots, I$, define $T_{Gi}$ to be independent random variables taking the value $\overline{q}_i = \max(q_{i1}, q_{i2})$ with probability $\Gamma/(1 + \Gamma)$ and the value $\overline{q}_i = \min(q_{i1}, q_{i2})$ with probability $1/(1 + \Gamma)$. Under Assumption 2 with $\Gamma \geq 1$, we have for any $t$,

$$\frac{P(T \geq t|X,M)}{P(T_{\text{exact}} \geq t)} \leq \frac{P_{T_{\text{approx}}}}{P_{T_{\text{exact}}}} \leq \Phi\left(\frac{t - \sum_{i=1}^I (\frac{\Gamma}{1+\Gamma} \overline{q}_i + \frac{1}{1+\Gamma} \overline{q}_i)}{\sqrt{\sum_{i=1}^I (\frac{\Gamma}{1+\Gamma} \overline{q}_i - \overline{q}_i)^2}}\right),$$

where

$$T_{\text{approx}} = \sum_{i<j \in \mathcal{T}^{(2)}} Z_{ij}g_{ij}(\hat{f}_{1-2}).$$
where \( \Phi(\cdot) \) is the distribution function of standard normal distribution, and "\( \simeq \)" denotes that two sequences are asymptotically equal as \( I \to \infty \).

**Proof.** All proofs in the article can be found in Appendix B. 

In Proposition 1, let \( I = |\mathcal{I}(T)| \) and \( q_{ij} = g_{ij}(\hat{f}_{1 \rightarrow 2}) \), and we can then calculate the bounding p-value \( p_{1 \rightarrow 2,\Gamma} \) under the biased RA restricted to \( i \in \mathcal{I}(T) \) for a fixed \( \Gamma \) value. This bounding p-value can be obtained by calculating the tail probability of the random variable \( \sum_{i \in \mathcal{I}(T)} \overline{q}_{ij} \) via Monte Carlo or via the Normal approximation. Next, flip the role of \( \mathcal{I}(T) \) and \( \mathcal{I}(Z) \), form a second test statistic \( T_{2 \rightarrow 1} = \sum_{i \in \mathcal{I}(Z)} \sum_{j=1}^{2} Z_{ij} h_{ij}(\hat{f}_{2 \rightarrow 1}) \) where \( h_{ij} : [0,1]^{\mathcal{I}(Z)} \to \mathbb{R} \) is another generic function, and obtain \( p_{2 \rightarrow 1,\Gamma} \) in a similar way. Finally, we reject Assumption 2 with a prespecified \( \Gamma \) at the level \( \alpha \) when \( \mathcal{P}_{T} = \min\{p_{1 \rightarrow 2,\Gamma}, p_{2 \rightarrow 1,\Gamma}\} < \alpha/2 \). Algorithm 1 summarizes this sample-splitting variant of the CPT, which we refer to as SS-CPT. Sample-splitting is an essential element of Algorithm 1 because scores \( q_{ij} \) in Proposition 1 are held fixed by sample-splitting; in contrast, scores in the vanilla version of the CPT reviewed in Section 1.3 depend on the permuted treatment labels and change at each permutation. It is unclear how to derive a bounding p-value and test the biased RA under the vanilla CPT.

What are some sensible choices of \( g_{ij} \) (and similarly \( h_{ij} \)) in Algorithm 1? There are at least three choices. First, we may let \( g_{ij}(\hat{f}_{1 \rightarrow 2}) = 1(\hat{f}(x_{ij}) > \hat{f}(x_{ij}')) \), and assign the predicted “treated” label to the unit with the higher predicted score and “control” label to the other within each matched pair. In this way, the test statistic \( T \) is essentially a prediction accuracy measure described in Gagnon-Bartsch and Shem-Tov (2019). Second, we may directly take the predicted score as \( q_{ij} \) by setting \( g_{ij}(\hat{f}_{1 \rightarrow 2}) = \hat{f}(x_{ij}) \). Third, we may take the rank of the predicted score for the unit \( ij \) among the predicted scores of all study units as \( q_{ij} \) by setting \( g_{ij}(\hat{f}_{1 \rightarrow 2}) = \sum_{i' = 1}^{I} \sum_{j' = 1}^{2} 1(\hat{f}(x_{ij}) \geq \hat{f}(x_{i'j'})) \).

### 4. Clustering-Based Test (CBT)

#### 4.1. A Clustering-Based Framework

If a clustering algorithm ALG can correctly group matched-pair data into a treated cluster and a control cluster solely based on the observed covariates, then this provides evidence that data is not well-matched or well-overlapped, and some degree of deviation from the RA persists. This intuition is illustrated in Figure 1. Definition 1 states that an algorithm ALGs is appropriate for testing Assumption 2 if it leverages no more information than what equation (2) conditions upon, and outputs “educated guesses” of cluster membership that respect the matched-pair structure. In particular, Definition 1 rules out using the treatment status (up to \( Z_{i1} + Z_{i2} = 1 \)) or outcome data.

**Definition 1 (Appropriateness).** Let \( \mathcal{F} = \{X, Z_{i1} + Z_{i2} = 1, \forall i = 1, \ldots, I\} \) be the collection of information. An algorithm ALG is appropriate if it takes as input \( \mathcal{F} \) and outputs a partition of \( 2I \) matched units into two equal-sized groups (each with size \( I \)): \( \Pi_{1} = \{ij : i = 1, \ldots, I\} \) and \( \Pi_{2} = \{ij : i = 1, \ldots, I\} \), subject to the constraint that there is one and only one treated unit in each matched pair \( i, \) that is, we have either \( i1 \in \Pi_{1} \) and \( i2 \in \Pi_{2} \) or \( i2 \in \Pi_{1} \) and \( i1 \in \Pi_{2} \).

Let ALG be an appropriate algorithm satisfying Definition 1. Proposition 2 specifies the null distribution of the number of “correctly guessed” cluster membership returned by ALG under the biased RA for a fixed \( \Gamma \). The null distribution under the RA is a special case of Proposition 2 by setting \( \Gamma = 1 \) and presented as Corollary 1 for easy reference.

**Proposition 2 (Bounding Two-Sided p-value).** Let \( \mathcal{T} = \{ij : Z_{ij} = 1, i = 1, \ldots, I, j = 1, 2\} \) be the set of indices corresponding to the treated units in each matched pair. Let \( \Pi_{1} \) and \( \Pi_{2} \) be the output from an appropriate algorithm ALG. For \( i = 1, \ldots, I \), define \( \overline{q}_{T_{i}} \) to be independent random variables taking the value 1 with probability \( \Gamma/(1 + \Gamma) \) and the value 0 with probability \( 1/(1 + \Gamma) \). Under Assumption 2 with \( \Gamma \geq 1 \), we have for any \( t \),

\[
P \left( \left| \sum_{i=1}^{I} \overline{q}_{T_{i}} \right| - I/2 \right)^{2} \geq (t - I/2)^{2} |X, M| \leq P \left( \sum_{i=1}^{I} \overline{q}_{T_{i}} \geq |t - I/2| + I/2 \right) + P \left( \sum_{i=1}^{I} (1 - \overline{q}_{T_{i}}) \leq -|t - I/2| + I/2 \right) \overset{\Delta}{=} P_{\text{exact}} \]

\[
\simeq 1 - \Phi \left( \frac{|t - I/2| + I/2 - (\Gamma/(1 + \Gamma)) \cdot I}{\Gamma/(1 + \Gamma)^{2} \cdot I} \right)^{1/2} \]

#### Algorithm 1: Pseudo Algorithm for Testing Assumption 2 using SS-CPT

**Input:** \( X = \{x_{ij} : i = 1, \ldots, I, j = 1, 2\} \) and \( Z = \{Z_{ij} : i = 1, \ldots, I, j = 1, 2\} \).

1. Randomly split the matched sample into two parts: \( \mathcal{I}(T) \) and \( \mathcal{I}(Z) \) such that \( \mathcal{I}(T) \cup \mathcal{I}(Z) = \{1, \ldots, I\} \) and \( \mathcal{I}(T) \cap \mathcal{I}(Z) = \emptyset \).

2. Train a classifier \( \hat{f}_{1} \) with covariates \( X^{(1)} = \{X_{ij} : i \in \mathcal{I}(T), j = 1, 2\} \) as predictors and treatment status \( Z^{(1)} = \{Z_{ij} : i \in \mathcal{I}(T), j = 1, 2\} \) as labels.

3. Train a classifier \( \hat{f}_{2} \) with covariates \( X^{(2)} = \{X_{ij} : i \in \mathcal{I}(Z), j = 1, 2\} \) as predictors and treatment status \( Z^{(2)} = \{Z_{ij} : i \in \mathcal{I}(Z), j = 1, 2\} \) as labels.

4. Let \( \hat{f}_{1 \rightarrow 2} = (\hat{f}_{1}(x_{ij}), i \in \mathcal{I}(T), j = 1, 2) \) denote the predicted scores for \( T^{(2)} \) based on the classifier \( \hat{f}_{1} \). Define the test statistic \( T_{1 \rightarrow 2} = \sum_{i \in \mathcal{I}(T)} \sum_{j=1}^{2} Z_{ij} h_{ij}(\hat{f}_{1 \rightarrow 2}) \), and calculate the worst-case p-value under the biased RA with \( \Gamma \) according to Proposition 1. Denote this bounding p-value by \( p_{1 \rightarrow 2,\Gamma} \).

5. Let \( \hat{f}_{2 \rightarrow 1} = (\hat{f}_{2}(x_{ij}), i \in \mathcal{I}(Z), j = 1, 2) \) denote the predicted scores for \( T^{(1)} \) based on the classifier \( \hat{f}_{2} \). Define the test statistic \( T_{2 \rightarrow 1} = \sum_{i \in \mathcal{I}(Z)} \sum_{j=1}^{2} Z_{ij} h_{ij}(\hat{f}_{2 \rightarrow 1}) \), and calculate the worst-case p-value under the biased RA with \( \Gamma \) according to Proposition 1. Denote this bounding p-value by \( p_{2 \rightarrow 1,\Gamma} \).

**Output:** Reject Assumption 2 with the prespecified \( \Gamma \) if and only if \( \min\{p_{1 \rightarrow 2,\Gamma}, p_{2 \rightarrow 1,\Gamma}\} < \alpha/2 \).
Figure 1. A schematic plot with two-dimensional observed covariates \((X_1, X_2)\). Top left panel: 20 treated and 20 control subjects after matching. Top right panel: the same data with treatment labels anonymized. This is the input into the clustering algorithm. Bottom panel: anonymized data partitioned into two groups (Group 1 versus Group 2) based on a clustering algorithm. This partition is then compared to the true labels (top left panel) and high clustering accuracy is used as evidence testing the RA and biased RA.

\[
\Delta = \frac{\Phi \left( -\frac{|I - I/2| + I/2 - (1/(1 + \Gamma)) \cdot I}{\Gamma/(1 + \Gamma)^2 \cdot I^{1/2}} \right)}{p_{\Gamma, \text{approx}}},
\]

**Corollary 1.** Using notation in Proposition 2 and under Assumption 1, we have \(\left| \Pi_1 \cap \mathcal{T} \right| \sim \text{Binomial} \left( I, 1/2 \right)\).

According to Proposition 2, the bias RA can be tested by comparing the number of "correctly inferred" cluster membership to a Binomial distribution and calculating appropriate tail probabilities. Both exact \(p\)-value \((p_{\Gamma, \text{exact}})\) and that based on normal approximation \((p_{\Gamma, \text{approx}})\) can be obtained efficiently.

### 4.2. Implementation of CBT: Clustering with Side Information

We propose two simple algorithms for CBT, one based on a variant of the 2-means clustering algorithm and the other based on fitting a two-component mixture model. Both algorithms perform the clustering task with the following side information: one and only one unit in each matched pair is the treated unit.

**Side-Information 1 (Matched-Set-Structure Side-Information).** Let \(\mathcal{I} = \{1, \ldots, I\}\) index the matched pairs, then treatment assignment indicators satisfy \(Z_{i1} + Z_{i2} = 1, \forall i \in \mathcal{I}\). Write \(\mathcal{F}_{\text{side,match}} = \{Z_{i1} + Z_{i2} = 1, \forall i \in \mathcal{I}\}\) and \(\mathcal{F}_{\text{side,match}}\) will be referred to as the matched-set-structure side-information.

Side-information 1 imposes the so-called Cannot-Link constraints to the conventional 2-means clustering algorithm when updating the cluster membership at each iteration, and is known as a constrained 2-means clustering algorithm in the literature (Wagstaff et al. 2001). In Appendix C, we describe in detail how to solve the constrained 2-means problem, and how to use a machine learning technique called "metric learning" to update the distance metric at each iteration of the algorithm, by (i) maximizing the distance between dissimilar pairs, that is, units \(i1\) and \(i2\) within each matched pair \(i\), and (ii) enforcing the distance between units in each cluster and the cluster centroid to be small.

In addition to \(K\)-means clustering, another popular clustering method in practice is fitting a mixture model. In our application, the observed covariates data naturally arises from a two-component mixture model \(\alpha_{treat} \cdot F_t(\cdot) + (1 - \alpha_{treat}) \cdot F_t(\cdot),\) where \(F_t(\cdot)\) and \(F_t(\cdot)\) represent some parametric family of covariates.
distribution in the treated and matched control groups, respectively. Moreover, the mixing probability \( \alpha_{\text{treat}} = 0.5 \) for the matched-pair design. After fitting a mixture model, say using the EM algorithm (Dempster, Laird, and Rubin 1977), one may then compare the relative magnitude of posterior probabilities of belonging to each cluster of two units in each matched pair, and assign cluster membership accordingly.

5. Quantifying Observed Covariates’ Residual Confounding in Matched Samples

The bounding \( p \)-values \( \{\tilde{p}_{\Gamma,\text{exact}}, \Gamma \geq 1\} \) testing the biased RA give rise to the following measure of residual confounding, termed residual sensitivity value (RSV), in a matched sample.

**Definition 2** (Residual Sensitivity Value \( \tilde{\Gamma} \)). Given matched-pair data, an exact test for Assumption 2, and a significance level \( \alpha \), the residual sensitivity value \( \tilde{\Gamma} \) is the smallest \( \Gamma \geq 1 \) such that the bounding \( p \)-value \( \tilde{p}_{\Gamma,\text{exact}} \) is not significant, that is,

\[
\tilde{\Gamma} := \inf\{\Gamma \geq 1\} \text{subject to} \tilde{p}_{\Gamma,\text{exact}} \geq \alpha.
\]  

The residual sensitivity value \( \tilde{\Gamma} \) is a measure of residual confounding from observed covariates in the matched-pair data. Any valid test of Assumption 2 can in principle be inverted sequentially to define the RSV, though we focus on the RSVs derived from SS-CPT and CBT in this article. It is an interesting research topic to explore other powerful statistical tests for Assumption 2.

By definition, if the RA cannot be rejected at the level \( \alpha \), then \( \tilde{p}_{\Gamma=1,\text{exact}} \geq \alpha \) and hence \( \tilde{\Gamma} = 1 \). The RSV \( \tilde{\Gamma} \) serves at least two practical purposes. First, a large \( \tilde{\Gamma} \) signals poor balance after statistical matching and empirical researchers may consider employing some useful study design strategies to improve their matched samples (Rosenbaum, Ross, and Silber 2007; Rosenbaum 2012; Pimentel et al. 2015; Zhang et al. 2021b). We elaborate on a few strategies in Appendix D and illustrate a strategy called optimal subset match (Rosenbaum 2012) when examining the case study in Section 7. Second, residual confounding from observed covariates needs to be taken into account in the downstream outcome analysis; the randomization-based outcome analysis should relax the RA and report a bounding \( p \)-value corresponding to taking \( \Gamma = \tilde{\Gamma} \) in a biased randomization scheme. We do need to point out again that although \( \tilde{\Gamma} \) provides a measure of residual confounding due to imperfect matching on observed covariates, which represents the smallest degree of deviation from the RA, a sensitivity analysis is still much needed to examine the possibility of unmeasured confounding. Researchers should not confuse the proposed RSV with Zhao’s (2018) sensitivity value, which measures the minimum strength of unmeasured confounding needed to qualitatively alter the outcome analysis.

6. Simulation Studies

6.1. Goal and Structure

We have four goals in the simulation section. First, we compared the power of two implementations of SS-CPT and two implementations of CBT. Second, we compared several statistical matching algorithms and investigated which algorithm produced matched samples with minimal residual imbalance as quantified by the RSV. Third, we examined the relationship between RSV and informal balance diagnostics like the SMDs. Lastly, we investigated the relationship between the performance of downstream randomization inference and both formal and informal diagnostics.

Our simulation can be summarized as a \( 2 \times 3 \times 3 \times 4 \) factorial design with the following factors:

**Factor 1**: sample size before matching, \( n \): 3000 and 5000.

**Factor 2**: observed covariates distribution and overlap: \( X \sim \text{Multivariate Normal} (\mu, \Sigma) \), with \( \mu_{\text{0x1}} = (cZ, 0, \ldots, 0)^T \) and \( \Sigma = I_{10 \times 10} \) with \( c = 0.3, 0.5, \) and 0.7.

**Factor 3**: statistical matching algorithm, \( M \):

1. \( M_{\text{maha}} \): metric-based matching with the Mahalanobis distance;
2. \( M_{\text{pscore}} \): propensity score matching with no caliper;
3. \( M_{\text{opt}} \): optimal matching within a 0.2 SD propensity score caliper.

**Factor 4**: testing procedure: (i). CBTK-means: CBT based on a constrained 2-means algorithm; (ii). CBTMM: CBT based on a two-component Gaussian mixture model; (iii). SS-CPT: CBT using \( g_{\hat{f}_{1\rightarrow 2}} (\hat{f}_{2\rightarrow 1}) = h_1 (\hat{f}_{2\rightarrow 1}) = 1 (\hat{f}_i (x_j) > \hat{f}_i (x_j')) \) as scores; (iv). SS-CPTpscore: SS-CPT using \( g_{\hat{f}_{1\rightarrow 2}} (\hat{f}_{2\rightarrow 1}) = h_1 (\hat{f}_{2\rightarrow 1}) = \hat{f}_i (x_j) \) as scores.

Factor 1 through 3 define the data-generating processes under consideration; see Rubin (1979) and Zhang et al. (2021b) for some motivation for this simulation setup. In particular, Factor 2 specifies the observed covariates distribution in the treated and control groups. Parameter \( c \) controls the amount of overlap in the treated and control groups before matching; \( c = 0.5 \) is often considered a moderately large bias (Rubin 1979, expression 3.3). We generated \( Z \sim \text{Binomial(1/3)} \) so that the treated-to-control ratio in the simulated datasets was approximately 1 to 2. Factor 3 defines three statistical matching algorithms under consideration. All matching algorithms were implemented using the R package `match2C` (Zhang et al. 2021b; Zhang 2021). A tutorial of the package can be found via `https://cran.r-project.org/web/packages/match2C/vignettes/tutorial.html`. Factor 4 represents four testing procedures to be studied. There are many other possible implementations of SS-CPT and CBT. Combining SS-CPT and CBT with more powerful machine learning methods may further improve their power. We considered these specific implementations because they are familiar to empirical researchers and easy-to-implement.

Lastly, we generated the potential outcomes \( R_T \) and \( R_C \) for each unit with selected \( k_1 \) and \( k_2 \) values as follows:

\[
R_T = R_C = k_2 \times \text{sign}(X_1) \cdot |X_1|^{k_1} + 0.5\sqrt{|X_2|} - X_3 + \epsilon,
\]

\[
\epsilon \sim \text{Normal}(0, 1).
\]

The observed outcome is \( R = R_T \cdot Z + R_C \cdot (1 - Z) \).

6.2. Measures of Success

For each data-generating process of \( (X, Z) \) and statistical matching algorithm defined by Factor 1 through 3, we repeated the
simulation 500 times and computed the proportion of times the RA was rejected by each of the four algorithms described in Factor 4. These rejection proportions are denoted by $\text{Power}_{\Gamma=1}$. We then calculated the residual sensitivity values $\Gamma$ determined by each of the four testing procedures for each matched sample. The RSVs complement the rejection proportions at $\Gamma = 1$ by quantifying the extent of deviation from the RA and reflecting the power of each procedure at $\Gamma > 1$. The average RSV is denoted as Mean $\Gamma$.

To facilitate our understanding of the relationship between informal balance diagnostics and formal statistical tests like those developed in this article, we also recorded and reported the average standardized mean difference (defined as the difference in means of a covariate in the treated and matched control group, divided by the pooled standard deviation before matching) of the first covariate $X_1$, denoted as $\text{SMD}_{X_1}$, and the average median SMD, denoted as $\text{SMD}_{0.50}$, across all 10 covariates. According to our data-generating process, the propensity score after matching.

Table 1 summarizes the simulation results for different choices of sample size $n$, overlap parameter $c$, and statistical matching algorithm $\mathcal{M}$, and testing procedure when $k_1 = k_2 = 0.2$ in the outcome model. We identified several consistent trends from simulation results. First, each of the four testing procedures had improved power when testing the RA ($\text{Power}_{\Gamma=1}$) and identified a larger RSV (Mean $\Gamma$) for larger sample size $n$ and worse matching quality as captured by $\text{SMD}_{X_1}$. The testing procedure SS-CPT pscore (Method IV in Table 1) seemed to have the largest power when testing the RA and identify the largest RSV in 15/18 simulation settings. Interestingly, SS-CPT pscore appeared to be superior to SS-CPT accuracy in testing the RA and identifying the RSV in almost all simulation settings. The testing procedure CBTGMM appeared to be slightly more favorable in the other 3 simulation settings where the sample was well matched and $\text{SMD}_{X_1}$ was small. Based on the simulation results, we would recommend using CBTGMM when the largest SMD is less than 0.05, or one-twentieth of one pooled standard deviation, and SS-CPT pscore otherwise.

Second, upon examining Table 1, we found that optimal $\mathcal{M}_{\text{opt}}$ delivered the best matched samples in the following senses: (i) the average SMD of $X_1$ which served as an

| $n$ | $\mathcal{M}$ | $\text{SMD}_{X_1}$ | $\text{SMD}_{0.50}$ | $\text{Power}_{\Gamma=1}$ | Mean $\Gamma$ | H-L est | RMSE |
|-----|---------------|--------------------|--------------------|--------------------------|-------------|--------|------|
|     |               |                    |                    |                          |             |        |      |
| 3000| $\mathcal{M}_{\text{opt}}$ | 0.01               | 0.01               | 0.04                     | 0.14        | 0.01   | 0.00 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.12               | 0.02               | 0.12                     | 0.31        | 0.93   | 0.97 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.30               | 0.04               | 0.36                     | 0.40        | 0.94   | 0.98 |
| 5000| $\mathcal{M}_{\text{opt}}$ | 0.01               | 0.01               | 0.05                     | 0.24        | 0.01   | 0.01 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.11               | 0.01               | 0.16                     | 0.45        | 0.99   | 1.00 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.30               | 0.03               | 0.56                     | 0.53        | 1.00   | 1.00 |
| 3000| $\mathcal{M}_{\text{opt}}$ | 0.05               | 0.02               | 0.06                     | 0.23        | 0.07   | 0.16 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.21               | 0.02               | 0.46                     | 0.47        | 1.00   | 1.00 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.50               | 0.04               | 0.82                     | 0.70        | 1.00   | 1.00 |
| 5000| $\mathcal{M}_{\text{opt}}$ | 0.05               | 0.01               | 0.05                     | 0.27        | 0.16   | 0.31 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.19               | 0.01               | 0.53                     | 0.56        | 1.00   | 1.00 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.50               | 0.03               | 0.92                     | 0.83        | 1.00   | 1.00 |
|     | $\mathcal{M}_{\text{opt}}$ | 0.14               | 0.02               | 0.06                     | 0.27        | 0.73   | 0.94 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.31               | 0.02               | 0.90                     | 0.64        | 1.00   | 1.00 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.70               | 0.03               | 0.99                     | 0.89        | 1.00   | 1.00 |
| 5000| $\mathcal{M}_{\text{opt}}$ | 0.13               | 0.01               | 0.04                     | 0.29        | 0.94   | 1.00 |
|     | $\mathcal{M}_{\text{maha}}$ | 0.29               | 0.01               | 0.95                     | 0.66        | 1.00   | 1.00 |
|     | $\mathcal{M}_{\text{pscore}}$ | 0.70               | 0.03               | 0.99                     | 0.95        | 1.00   | 1.00 |

NOTE: We let $k_1 = k_2 = 0.2$ in the outcome model. Roman numerals I to IV are shorthands for CBTKM (I), CBTGMM (II), SS-CPT accuracy (III), and SS-CPT pscore (IV). The highest power at $\Gamma = 1$ and largest mean RSV for each data-generating process and matching algorithm are highlighted. Bold values are highlighted for highest power and largest RSV.
informal measure of overall balance was much smaller for $M_{opt}$ compared to $M_{pscore}$ or $M_{maha}$; (ii) with even a moderately large bias before matching ($c = 0.5$) and a large sample size ($n = 5000$), testing procedures could barely reject the RA on an optimally matched sample (highest Power$_{1} = 0.31$ for SS-CPT$_{pscore}$), and identify little deviation (largest mean RSV $= 1.03$ for SS-CPT$_{pscore}$).

Third, we concluded that RSVs supplemented informal diagnostics like SMDs. In empirical studies, an informal rule of thumb says that SMDs of all observed covariates should be less than 0.1 (Silver et al. 2001; Austin and Stuart 2015). In our simulation studies, many matched samples satisfied this rule of thumb, for example, optimal matching when $c = 0.3$ or 0.5 and Mahalanobis-distance-based matching when $c = 0.3$; however, even in these circumstances, the RA was frequently rejected. For instance, when $c = 0.3$ and $M = M_{maha}$, SS-CPT$_{pscore}$ rejected 485/500 matched datasets and identified an RSV as large as 1.31 on average. The bottom line is that the RSVs provide a formal and quantitative assessment of the quality of matched sample. We also repeated a subset of simulation studies with multiple random initializations when running CBT and multiple random sample-splitting schemes when running SS-CPT, and found that Power$_{1} = 1$ and Mean $\hat{\beta}$ remained stable.

What are the implications of statistical matching quality on the randomization-based outcome analysis? Table 1 suggests that the root mean squared error (RMSE) of the randomization-based outcome analysis of optimally-matched samples ($M = M_{opt}$) seemed to be the best among three matching methods across all data-generating processes. Figure 2 further presents the boxplots of the Hodges-Lehmann point estimate in matched samples when the sample size $n = 3000$ and the RSV as determined by SS-CPT$_{pscore}$ equal to 1 (i.e., the RA is not rejected) versus those with RSV larger than 1 (i.e., the RA is rejected). We found that the estimate was less biased when the RA was not rejected by SS-CPT$_{pscore}$.

Table 2 summarizes the bias and coverage probability of an outcome analysis assuming randomization (i.e., $\Gamma = 1$), and an outcome analysis that takes into account the RSV $\hat{\Gamma}$. obtained using SS-CPT$_{pscore}$ and conducts the outcome analysis under $\Gamma = \hat{\Gamma}$ when $n = 3000$, $c = 0.3$, and different $k_1$ and $k_2$ combinations. We found several consistent trends. First, the testing procedure SS-CPT$_{pscore}$ has little power against optimally-matched datasets in this simulation setting (i.e., the RSV $\hat{\Gamma}$ is only occasionally greater than 1; see Table 1), and the coverage probabilities under both $\Gamma = 1$ and $\Gamma = \hat{\Gamma}$ are similar and close to the nominal rate. Second, $M_{maha}$ and $M_{pscore}$ tended to remove less bias and as a result, outcome analyses have poor coverage probabilities under $\Gamma = 1$; however, the partial identification intervals obtained under $\Gamma = \hat{\Gamma}$ almost always contain the true estimand $\beta = 0$ and the bias is negligible. Moreover, the confidence intervals corresponding to $\Gamma = \hat{\Gamma}$ always obtain the nominal level, although conservative, due to the conservative nature of partial identification and Rosenbaum bounds (i.e., Rosenbaum bounds is a worst-case $p$-value).

| $k_2$ | $k_1$ | $M$  | 100× Bias $\Gamma = 1$ | 100× Bias $\Gamma \neq 1$ | Coverage $\Gamma = 1$ | Coverage $\Gamma \neq 1$ |
|-------|-------|-----|------------------------|--------------------------|-------------------|-------------------|
| 0.20  | 0.20  | $M_{opt}$ | 2.06 | 2.03 | 94.6% | 94.6% |
| 0.10  | 0.00  | $M_{maha}$ | 2.70 | 0.15 | 89.6% | 99.8% |
| 0.10  | 0.00  | $M_{pscore}$ | 2.99 | 0.04 | 93.4% | 100% |

NOTE: Sample size $N = 3000$, overlap parameter $c = 0.3$, and the RSV $\hat{\Gamma}$ obtained by SS-CPT$_{pscore}$.
7. Case Study: Effectiveness of Right Heart Catheterization (RHC) in the Initial Care of Critically Ill Patients

7.1. Background, All Samples and Matched Samples

We illustrate various aspects of our proposed framework by applying it to an observational study investigating the effectiveness of right heart catheterization (RHC) in the initial care of critically ill patients. Since its introduction into the intensive care units (ICUs) almost 50 years ago, RHC was perceived by clinicians as largely beneficial because direct measurements of heart function by RHC help guide therapies, which was believed to lead eventually to more favorable patient outcomes (Connors, McCaffree, and Gray 1983; Connors et al. 1996). In fact, the belief in RHC’s effectiveness was so strong that many physicians refused to enroll their patients into randomized controlled trials for ethical considerations (Guyatt 1991). In the absence of RCTs, Connors et al. (1996) examined the effectiveness of RHC in a matching-based cohort study using observational data. We followed Rosenbaum (2012) and considered patients under the age of 65 from the work of Connors et al. (1996); among these patients are 1194 who received RHC and 1804 who did not. We followed Connors et al. (1996) and considered observed covariates related to patients demographics, laboratory measurements and vital signs because these factors clearly affect both physicians’ decision to use or withhold RHC and patients’ outcomes. The outcome of interest is patients’ 30-day mortality.

The first three columns of Table 3 summarize patients’ characteristics in the full sample. Not surprisingly, a number of important variables are vastly different between the two groups; for instance, the mean blood pressure is only 69.44 mm Hg in the RHC group compared to 87.47 mm Hg in the no RHC group. Overall, the two groups before matching have similar characteristics in the full sample. Not surprisingly, a number of important variables are vastly different between the two groups; for instance, the mean blood pressure is only 69.44 mm Hg in the RHC group compared to 87.47 mm Hg in the no RHC group. Upon applying SS-CPTaccuracy and SS-CPTpscore to the design M1, we obtained an RSV of 2.52 and 5.67, respectively. The CBT based on fitting a two-component Gaussian mixture model yielded an RSV equal to 1.21. All three tests rejected the RA for M1, though two SS-CPT implementations were more powerful, which seemed to agree with our simulation results in similar settings. Next, we performed a randomization-based outcome analysis using McNemar’s test against the alternative hypothesis that RHC has a negative effect on 30-day mortality. Under the randomization assumption, the p-value of outcome analysis was 0.022, which provided some weak evidence against the null hypothesis of no effect. However, this treatment effect immediately disappeared when taking into account the RSVs obtained using either SS-CPT or CBT. In other words, the observed treatment effect under the RA may be an artifact of imperfect matching and residual imbalance in observed covariates, and a researcher who did not fully take this into account could draw a false conclusion.

On the other hand, none of our proposed tests rejected the RA for the design M2. Again, this aligns well with our simulation results as all absolute SMDs are less than 0.05 in M2, and we found in simulation studies that our testing procedures had little power in similar settings. For the matched design M2, we conducted outcome analysis under the RA, and obtained a p-value equal to 0.79, suggesting insufficient evidence against the causal null hypothesis for the study units involved in the design M2. All outcome analyses (under \( \Gamma = 1 \) or \( \Gamma = \tilde{\Gamma} \)) were performed using R package rcbsubset (Keele 2014).

Table 3. Balance table of all samples and the matched samples.

|                | All RHC (n = 1194) | All No RHC (n = 1804) | Abs. SMD | Optimal RHC (n = 1194) | Optimal No RHC (n = 1194) | Abs. SMD | Subset RHC (n = 1061) | Subset No RHC (n = 1061) | Abs. SMD |
|----------------|-------------------|-----------------------|----------|------------------------|---------------------------|----------|---------------------|---------------------------|----------|
| **Demographics** |                   |                       |          |                        |                           |          |                     |                           |          |
| Age, yr        | 49.56             | 48.01                 | 0.09     | 49.26                  | 48.65                     | 0.02     | 48.65               | 48.07                     | 0.03     |
| Male, %        | 0.58              | 0.57                  | 0.02     | 0.59                   | 0.57                      | 0.01     | 0.57                | 0.58                      | 0.02     |
| White, %       | 0.72              | 0.72                  | 0.01     | 0.73                   | 0.73                      | 0.02     | 0.73                | 0.72                      | 0.01     |
| **Laboratory Measurements** |     |                       |          |                        |                           |          |                     |                           |          |
| PaO2/FiO2, mm Hg | 196.13            | 243.68                | 0.29     | 214.55                 | 225.81                    | 0.11     | 226.36              | 226.36                    | 0.00     |
| PaO2, mm Hg     | 36.74             | 38.67                 | 0.11     | 37.62                  | 37.55                     | 0.05     | 37.38               | 0.01                      |          |
| WBC count, ×10⁹/L | 15.77             | 14.98                 | 0.05     | 15.29                  | 15.84                     | 0.03     | 15.30               | 0.03                      |          |
| Creatinine, μmol/L | 2.46              | 1.94                  | 0.17     | 2.15                   | 2.07                      | 0.10     | 2.16                | 0.03                      |          |
| **Vital signs** |                   |                       |          |                        |                           |          |                     |                           |          |
| Blood pressure, mm Hg | 69.44             | 87.47                 | 0.35     | 75.95                  | 80.07                     | 0.13     | 79.86               | 0.00                      |          |
| APACHE II score | 61.07             | 50.51                 | 0.36     | 55.63                  | 53.34                     | 0.19     | 54.56               | 0.04                      |          |
| Coma score      | 17.90             | 21.66                 | 0.09     | 16.49                  | 14.69                     | 0.03     | 20.40               | 20.03                     | 0.01     |
| Propensity score | 0.48              | 0.35                  | 0.55     | 0.42                   | 0.24                      | 0.40     | 0.40                | 0.00                      |          |

NOTE: There are a total of 1194 patients under 65 undergoing RHC and 1804 not undergoing RHC. A total of 1194 matched pairs were formed in the optimal match M1 and 1061 pairs were formed in the optimal subset match M2. SMD is the abbreviation of the standardized mean difference which is equal to the difference in means of covariates in the treated and matched control groups divided by the pooled standard deviation in the treated and control groups before matching.
In this example, there is a tension between the internal validity (i.e., comparability of the RHC and no RHC groups) and external validity (i.e., how results generalize to a target population): the design M1 has superior generalizability over M2 as M1 uses the entire treated group, while the design M2 has far superior internal validity as it better approximates an ideal RCT; see Zhang (2022) for a method that builds a well-matched sample mimicking a target population.

8. Discussion: Summary, Strengths and Weaknesses of Proposed Tests, and Extensions

It is a popular strategy in comparative effectiveness research to embed observational data into a randomized controlled experiment using statistical matching and analyze matched data as if it were a randomized experiment. As Collin Mallows famously pointed out (see, e.g., Denby and Landwehr 2013), the most robust statistical technique is to look at the data; a matched observational study is therefore robust in the sense that it forces researchers to examine the covariate balance and overlap, focus on the covariate space, that is, well-overlapped, and avoid unfounded extrapolation (Ho et al. 2007; Rubin 2007; Stuart 2010; Rosenbaum 2002, 2010). Despite a preferred strategy to draw causal conclusion (in our opinion), there is a gap between an approximate experiment (i.e., data after statistical matching) and a genuine experiment, and this gap is often circumvented by making the randomization assumption justified by informal or formal balance diagnostics.

One important limitation of statistical tests developed for the randomization assumption is that these tests cannot quantify the extent to which the randomization assumption is violated due to residual imbalance in observed covariates. Our proposed testing framework is thus advantageous in its ability to quantify the deviation from the randomization assumption using the residual sensitivity value \( \Gamma \). Although our primary focus in the article is matched-pair design, the framework and algorithm can be readily extended to matching with multiple controls; see Appendix B for analogous results for matching-with-multiple-controls designs.

Both SS-CPT and CBT can be used in combination with any user-chosen classification or constraint clustering algorithms. Our simulation studies compared only two specific implementations of each method; more powerful classification and constraint clustering methods could potentially deliver more powerful statistical tests. The SS-CPT method with a propensity score defined score function appeared to be the most powerful in most settings, while the CBT method based on Gaussian mixture modeling appeared to be slightly more advantageous in very closely matched sample. Compared to that of the CBT method, calculation of the exact bounding \( p \)-value of the SS-CPT method is less computationally efficient for a generic score function.

We recommend empirical researchers to examine the covariate balance using both formal and informal diagnostics, and when possible, incorporate the level of residual confounding into their outcome analysis. For instance, one way to do this is to perform a randomization inference under a biased randomization scheme using Rosenbaum bounds (Rosenbaum 2002, 2010) with the parameter \( \Gamma \) set to the magnitude of the residual sensitivity value, that is, \( \Gamma = \Gamma \). Other strategies to formally reflect the study design quality in the downstream outcome analysis are worth exploring.

Failure to reject our proposed test, like failure to reject any statistical test, does not translate to a statement about the correctness of the randomization assumption; in fact, statistical matching algorithms are likely to create dependence among matched pairs or sets so that the independence part of the randomization assumption almost surely does not hold. However, through our extensive simulations (see also simulations in Brandon 2020), it appears that when the randomization assumption cannot be rejected by our proposed tests, the randomization-based outcome analysis typically has good statistical performance; in other words, the randomization assumption is a good approximation of the complicated reality in these cases.

Lastly, in order to draw high-quality, convincing causal conclusions, one necessarily needs to perform extensive sensitivity analysis that allows for some hypothetical unmeasured confounding. To stress, while our proposed residual sensitivity value takes into account the deviation from the randomization assumption due to residual observed covariates imbalance, it says nothing about unmeasured confounding; in fact, unmeasured confounding can still bias the random assignment in each matched pair to an arbitrary extent even when the residual sensitivity value is 1. We recommend reporting the outcome analysis with both a residual sensitivity value and Zhao’s (2018) sensitivity value that examines the maximum extent of deviation from randomization (possibly due to unmeasured confounding) needed to qualitatively alter the causal conclusion are reported.

Supplementary Materials

Appendix contains additional literature review, extension of the proposed methodology to matching-with-multiple-controls, proofs, details on clustering algorithms, and practical strategies for improving a matched comparison. Code and data can be found in the code_and_data.zip file.

Acknowledgments

We would like to acknowledge the editor, associate editor, and three anonymous reviewers for their careful reviews and constructive comments which largely improved the article.

Funding

Research reported in this publication was partially supported by the National Institutes of Health (award RF1AG063481).

Conflict of Interest

The authors report there are no competing interests to declare.

ORCID

Qi Long © http://orcid.org/0000-0003-0660-5230
Bo Zhang © http://orcid.org/0000-0002-4381-1124
References

Austin, P. C., and Stuart, E. A. (2015), "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine*, 34, 3661–3679. [529,535]

Bind, M.-A. C., and Rubin, D. B. (2019), "Bridging Observational Studies and Randomized Experiments by Embedding the Former in the Latter." *Statistical Methods in Medical Research*, 28, 1958–1978. [528]

Branson, Z. (2020), "Randomization Tests to Assess Covariate Balance When Designing and Analyzing Matched Datasets." *Observational Studies*, 7, 1–36. [529,537]

Branson, Z., and Keele, L. (2020), "Evaluating a Key Instrumental Variable Assumption Using Randomization Tests." *American Journal of Epidemiology*, 189, 1412–1420. [529]

Connors, A. F., McCaffree, D. R., and Gray, B. A. (1983), "Evaluation of Right-Heart Catheterization in the Critically Ill Patient Without Acute Myocardial Infarction." *New England Journal of Medicine*, 308, 263–267. [536]

Connors, A. F., Steroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients." *JAMA*, 276, 889–897. [536]

Deheja, R. H., and Wahba, S. (2002), "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*, 84, 151–161. [529]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, Series B, 39, 1–22. [533]

Denby, L., and Landwehr, J. (2013), "A Conversation with Colin L. Mallows." *International Statistical Review*, 81, 338–360. [537]

Diamond, A., and Sekhon, J. S. (2013), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics*, 95, 932–945. [529]

DiPrete, T. A., and Gangl, M. (2004), "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology*, 34, 271–310. [528]

Fogarty, C. B. (2020), "Studentized Sensitivity Analysis for the Sample Average Treatment Effect in Paired Observational Studies." *Journal of the American Statistical Association*, 115, 1518–1530. [528]

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., and Schneeweiss, S. (2014), "Metrics for Covariate Balance in Cohort Studies of Causal Effects." *Statistics in Medicine*, 33, 1685–1699. [529]

Gagnon-Bartsch, J., and Shem-Tov, Y. (2019), "The Classification Permutation Test: A Flexible Approach to Testing for Covariate Imbalance in Observational Studies." *Annals of Applied Statistics*, 13, 1464–1483. [529,531]

Guyatt, G. (1991), "A Randomized Control Trial of Right-Heart Catheterization in Critically Ill Patients." *Journal of Intensive Care Medicine*, 6, 91–95. [536]

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014), "Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies." *Journal of the American Statistical Association*, 109, 133–144. [529]

Heller, R., Rosenbaum, P. R., and Small, D. S. (2010), "Using the Cross-Match Test to Appraise Covariate Balance in Matched Pairs." *The American Statistician*, 64, 299–309. [529]

Heng, S., Kang, H., Small, D. S., and Fogarty, C. B. (2021), "Increasing Power for Observational Studies of Aberrant Response: An Adaptive Approach." *Journal of the Royal Statistical Society*, Series B, 83, 482–504. [529]

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric causal Inference." *Political Analysis*, 15, 199–236. [537]

Iacus, S. M., King, G., and Porro, G. (2011), "Multivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association*, 106, 345–361. [529]

Keefe, L. J. (2014), *rbounds: Perform Rosenbaum Bounds Sensitivity Tests for Matched and Unmatched Data*. R package version 2.1. [536]

Pimentel, S. D. (2022), *rsubset: Optimal Subset Matching with Refined Covariate Balance*. R package version 1.1.7. [536]

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons." *Journal of the American Statistical Association*, 110, 515–527. [529]

Rosenbaum, P. R. (1989), "Optimal Matching for Observational Studies." *Journal of the American Statistical Association*, 84, 1024–1032. [529]

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association*, 102, 75–83. [533]

Rosenbaum, P. R., and Rubin, D. B. (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician*, 39, 33–38. [529]

Rubin, D. B. (1978), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association*, 74, 318–328. [533]

(2007), "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine*, 26, 20–36. [537]

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), "Multivariate Matching and Bias Reduction in the Surgical Outcomes Study." *Medical Care*, 39, 1048–1064. [529,535]

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, 25, 1–21. [528,537]

Wagstaff, K., Caride, C., Rogers, S., and Schroedl, S. (2001), "Constrained k-Means Clustering with Background Knowledge." In *Proceedings of the Eighteenth International Conference on Machine Learning*, Vol. 1, 577–584. [532]

Wu, I., and Ding, P. (2020), "Randomization Tests for Weak Null Hypotheses in Randomized Experiments." *Journal of the American Statistical Association*, 116, 1989–1913. [528]

Zhang, B. (2021), *match2C: Match One Sample using Two Criteria*. R package version 1.1.0. [533]

Zhang, B. (2022), "Towards Better Reconciling Randomized Controlled Trial and Observational Study Findings: Efficient Algorithms for Building Representable Matched Samples with Enhanced External Validity." [537]

Zhang, B., Heng, S., Mackay, E. J., and Ye, T. (2021a), "Bridging Preference-Based Instrumental Variable Studies and Cluster-Randomized Encouragement Experiments: Study Design, Noncompliance, and Average Cluster Effect Ratio." *Biometrics*. [529]

Zhang, B., Mackay, E. J., and Baisochi, M. (2022), "Statistical Matching and Subclassification with a Continuous Dose: Characterization, Algorithm, and Application to a Health Outcomes Study." *Annals of Applied Statistics*. [529]

Zhang, B., Small, D., Lasater, K., McHugh, M., Silber, J., and Rosenbaum, P. (2021b), "Matching One Sample According to Two Criteria in Observational Studies." *Journal of the American Statistical Association*, 1–34. [529,533]

Zhao, Q. (2018), "On Sensitivity Value of Pair-Matched Observational Studies." *Journal of the American Statistical Association*, 114, 713–722. [533,537]

Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery." *Journal of the American Statistical Association*, 107, 1360–1371. [529]