Observational Studies 7(2021) 1-36 Submitted 3/20; Published 3/24

Randomization Tests to Assess Covariate Balance When Designing and Analyzing Matched Datasets

Zach Branson **zach@stat.cmu.edu**
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Abstract

Causal analyses for observational studies are often complicated by covariate imbalances among treatment groups, and matching methodologies alleviate this complication by finding subsets of treatment groups that exhibit covariate balance. It is widely agreed upon that covariate balance can serve as evidence that a matched dataset approximates a randomized experiment, but what kind of experiment does a matched dataset approximate? In this work, we develop a randomization test for the hypothesis that a matched dataset approximates a particular experimental design, such as complete randomization, block randomization, or rerandomization. Our test can incorporate any experimental design, and it allows for a graphical display that puts several designs on the same univariate scale, thereby allowing researchers to pinpoint which design—if any—is most appropriate for a matched dataset. After researchers determine a plausible design, we recommend a randomization-based approach for analyzing the matched data, which can incorporate any design and treatment effect estimator. Through simulation, we find that our test can frequently detect violations of randomized assignment that harm inferential results. Furthermore, through simulation and a real application in political science, we find that matched datasets with high levels of covariate balance tend to approximate balance-constrained designs like rerandomization, and analyzing them as such can lead to precise causal analyses. However, assuming a precise design should be proceeded with caution, because it can harm inferential results if there are still substantial biases due to remaining imbalances after matching. Our approach is implemented in the `randChecks` R package, available on `CRAN`.

Keywords: covariate balance, experimental design, matching, randomization-based inference, randomization test, rerandomization

1. Introduction: Matching Approximates Randomized Experiments, But What Kind of Randomized Experiments?

Randomized experiments are often considered the "gold standard" of causal inference because they, on average, achieve balance on all covariates—both observed and unobserved—across treatment groups (Rubin, 2008a). However, in observational studies, pretreatment covariates often affect subjects' probability of receiving treatment. As a result, covariate distributions across treatment groups can be very different, leading to biased treatment effect estimates. Furthermore, treatment effect estimates will be more sensitive to model specification, because differing covariate distributions force models to extrapolate; thus,

statistical models that make covariate adjustments may still be unreliable (Rubin, 2007). Methods must be employed to address this systematic covariate imbalance.

One popular method is matching, where subjects in the treatment group are matched to similar subjects in the control group, and unmatched subjects are discarded, thereby producing a matched dataset of comparable treatment and control subjects. In this sense, matching can be viewed as a prepocessing step that produces a subset of the treatment and control groups that exhibits covariate balance (Ho et al., 2007). To obtain a subset that exhibits covariate balance, many matching algorithms explicitly pair or block treatment subjects to control subjects that have similar covariate values, as in propensity score matching (Rosenbaum & Rubin, 1985), $1 : k$ optimal matching (Rosenbaum, 1989; Ming & Rosenbaum, 2001; Lu et al., 2011), and coarsened exact matching (Iacus et al., 2011, 2012). Other recent matching algorithms optimize for covariate balance directly, as in covariate-balancing propensity scores (Imai & Ratkovic, 2014; Zhao, 2019; Vegetabile et al., 2019), kernel optimal matching (Kallus, 2020), and mixed integer programming approaches (Zubizarreta, 2012; Zubizarreta et al., 2014; Zubizarreta & Keele, 2017). The main motivation for matching is that it produces a dataset with covariate balance, thereby making treatment effect estimators less biased as well as less sensitive to model misspecification, because there is less need for models to extrapolate when estimating causal effects (Ho et al., 2007; Stuart, 2010; Abadie & Spiess, 2020).

The process of finding a subset that exhibits covariate balance is often called the design stage of an observational study, because the goal is to obtain a dataset whose treatment assignment mechanism approximates an experimental design (Rubin, 2007, 2008b; Rosenbaum, 2010). The assumption that a matched dataset approximates a randomized experiment is often justified by demonstrations of covariate balance between the treatment and control groups, as we would expect from a randomized experiment; however, how to best assess covariate balance for matched datasets is still an open problem. Common balance diagnostics are tables and graphical displays of standardized covariate mean differences (Ahmed et al., 2006; Stuart, 2010; Zubizarreta, 2012) and significance tests like t-tests and Kolmogorov-Smirnov tests (Lu et al., 2001; Bind & Rubin, 2019). For example, a rule-of-thumb is that standardized covariate mean differences of a matched dataset should be below 0.1 (Normand et al., 2001; Austin, 2009b; Zubizarreta, 2012; Resa & Zubizarreta, 2016). However, many recommend tighter covariate balance if possible: Stuart (2010) recommends choosing the matching algorithm "that yields the smallest standardized difference of means across the largest number of covariates," and Imai et al. (2008) recommends that "imbalance with respect to observed pretreatment covariates...should be minimized without limit where possible." Thus, it is common to run many matching algorithms until some prespecified covariate balance diagnostics are met, and then analyze the matched dataset as if it were from a randomized experiment (Dehejia & Wahba, 2002; Ho et al., 2007; Caliendo & Kopeinig, 2008; Harder et al., 2010).

It is widely agreed upon that covariate balance serves as evidence that a matched dataset approximates a randomized experiment, but what kind of randomized experiment does a matched dataset approximate? Importantly, the aforementioned diagnostics are not formal tests that a matched dataset approximates a randomized experiment; rather, they are rules-of-thumb. For example, because many matching algorithms pair or block treatment subjects with control subjects, it is common to analyze a matched dataset as if it were from a block

randomized experiment if covariate balance diagnostics are satisfied, even though these diagnostics do not test the hypothesis that treatment is randomized within blocks (Austin, 2008; Rubin, 2008b; Iacus et al., 2012). Because of this, it is an ongoing debate as to whether the blocked structure within matched datasets should be ignored when analyzing a matched dataset (Ho et al., 2007; Schafer & Kang, 2008; Stuart, 2010; Gagnon-Bartsch & Shem-Tov, 2019). The choice of assignment mechanism can have substantial implications on inferential results when analyzing a matched dataset, making this an important distinction to clarify. For example, it is well-known that block randomized experiments tend to yield more precise treatment effect estimation than completely randomized experiments (Box et al. 1978, Chapter 3, Greevy et al. 2004, Imbens & Rubin 2015, Chapter 4); because of this, King & Nielsen (2019) argued that it is preferable to use a matching algorithm designed to approximate a block randomized experiment (like coarsened exact matching) over a matching algorithm designed to approximate a completely randomized experiment (like propensity score matching). Similarly, recent works have found that rerandomized experiments—where subjects are randomized until covariate balance is achieved—yield more precise treatment effect estimation than completely randomized experiments (Li et al., 2018b; Li & Ding, 2020). Thus, it is arguably preferable to have a matched dataset that approximates a rerandomized experiment than a completely randomized experiment. Indeed, Stuart (2010) noted that the iterative approach of matching until covariate balance is achieved is similar to rerandomization, and thus a carefully designed matched dataset may indeed approximate a rerandomized experiment. However, current covariate balance diagnostics cannot ascertain what type of experimental design a matched dataset approximates, if any design at all.

In this work, we develop covariate balance diagnostics that do explicitly test the hypothesis that treatment is randomly assigned, thereby assessing whether or not a matched dataset approximates a completely randomized experiment, a block randomized experiment, or some other experimental design like rerandomization. All of our diagnostics follow a simple two-step procedure: Use a randomization test to quantify the distribution of covariate balance we would expect from a randomized experiment, and then determine if the observed covariate balance in the matched dataset is reasonably within that distribution. Other works have also proposed randomization tests for assessing covariate balance (Hansen, 2008; Hansen & Bowers, 2008; Cattaneo et al., 2015; Lee, 2013; Gagnon-Bartsch & Shem-Tov, 2019); our test is a generalization of these procedures, where any experimental design and any measure of covariate balance can be incorporated in the test.

There are four benefits to our procedure. First, our procedure validly tests the hypothesis that treatment is randomly assigned, in the sense that our α-level test controls the probability that we falsely reject this hypothesis. Second, our procedure can be easily combined with common diagnostics like Love plots (Ahmed et al., 2006)—which visualize standardized covariate mean differences—thereby providing researchers a data-driven way to assess if the observed covariate balance justifies assuming a particular experimental design for a matched dataset, instead of relying on rules-of-thumb that may vary from researcher to researcher. Third, our procedure allows researchers to assess any experimental design for a matched dataset, including rerandomization designs with covariate balance constraints. This suggests that well-designed matched datasets should perhaps be analyzed as if they are from well-designed rerandomized experiments, a notion we explore throughout this paper.

Finally, our procedure allows for a graphical display that puts several experimental designs on the same univariate scale, thereby allowing researchers to pinpoint which experimental design—if any—is most appropriate for a particular matched dataset. All of our tests and diagnostics are provided in the `randChecks` `R` package (available on `CRAN`), such that researchers can easily assess if a treatment is randomly assigned for any dataset.

Our procedure also has limitations, which stem from the fact that our procedure is a balance test—i.e., a procedure that tests for covariate balance. The limitations of balance tests have been widely discussed, to the point that some have recommended against balance tests entirely (Imai et al., 2008; Austin, 2008; Stuart, 2010). The most fundamental limitation of our procedure is that it only tests whether a particular design does *not* hold for a given matched dataset; if our test fails to reject the null hypothesis of random treatment assignment, it may be because the treatment is effectively randomized, or it may be because our test is underpowered in detecting violations of random assignment. This is not a limitation of our test specifically but of hypothesis testing in general. Nonetheless, balance tests are valuable in that they can detect clear violations of random assignment that would undermine causal inferences for matched data (Hansen, 2008; Lee, 2013). Two other critiques of balance tests for matched data—most famously discussed in Imai et al. (2008)—are that (1) they often refer to a hypothetical super-population when covariate balance is a characteristic of the matched sample, and (2) their statistical power is affected by sample size. Randomization tests address the first critique by computing an exact randomization distribution for the sample at hand, instead of utilizing asymptotic approximations. However, randomization tests do not address the second critique—indeed, our test will tend to fail to reject random assignment for very small samples due to low statistical power, but it is also the case that analyzing the corresponding matched dataset will have low power when estimating treatment effects (Hansen, 2008). We recommend that researchers quantify the magnitude of covariate imbalances alongside our randomization tests—as is done throughout this paper and in our `randChecks` `R` package—so that they can ascertain if covariate balance is still inadequate even if our test fails to reject random treatment assignment.

We recommend the following approach for designing and analyzing matched data. First, researchers should specify the type of experimental design (e.g., complete randomization or block randomization) they would like to approximate from the outset of the observational study. Then, researchers can match subjects in an attempt to approximate this design, where they can use our diagnostics in `randChecks` to assess if a particular design is plausible. We present these diagnostics in Section 2. After an experimental design is deemed plausible, it can be used to analyze the matched dataset; we outline how to do this in Section 3. In Section 4, we will find via simulation that well-designed matched datasets tend to approximate rerandomized experiments with covariate balance constraints, and analyzing matched datasets as such can yield more precise inference for causal effects. However, assuming a precise design should be proceeded with caution, because it can harm inferential results if there are still substantial biases due to covariate imbalances that remain after matching. In Section 5, we apply our approach to a causal analysis conducted in political science by Keele et al. (2017), who used subject-matter expertise to target balancing relevant covariates when constructing a matched dataset. We show how our covariate balance diagnostics can ascertain the type of experimental design their matched dataset approxi-

mates, as well as how we can condition on this covariate prioritization when analyzing the matched dataset. In Section 6, we conclude with discussions about extending our approach.

2. The Design Stage: Determining Plausible Experimental Designs for Matched Datasets

2.1 Setup, Notation, and Assumptions

Consider a matched dataset containing N subjects with an $N \times K$ covariate matrix $\mathbf{X}$ and binary treatment assignment $\mathbf{W} \equiv (W_1, \ldots, W_N)$, where $W_i = 1$ denotes subject i receiving treatment and $W_i = 0$ denotes control. Let $(Y_i(1), Y_i(0))$ denote subject i's potential outcomes under treatment and control, respectively.[1] Only one of the potential outcomes is observed for each subject, depending on the realization of $\mathbf{W}$. We follow the "Rubin causal model" (Holland, 1986) and assume that the covariates and potential outcomes are fixed, and thus only $\mathbf{W}$ is stochastic. The treatment and control subjects in the matched dataset may or may not be explicitly paired or blocked. We assume that the causal estimand is the average treatment effect (ATE), $\tau \equiv \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$, which may differ from the ATE in some larger, pre-matched dataset. Inference about the N subjects can be generalized to the extent that they are representative of a larger population.

Because only $\mathbf{W}$ is stochastic, it is essential to understand its distribution, which may depend on the potential outcomes $\mathbf{Y}(1) \equiv (Y_1(1), \ldots, Y_N(1))$ and $\mathbf{Y}(0) \equiv (Y_1(0), \ldots, Y_N(0))$, as well as on the covariates $\mathbf{X}$. However, neither $\mathbf{Y}(1)$ nor $\mathbf{Y}(0)$ are ever completely observed. We employ two assumptions that constrain the distribution of $\mathbf{W}$ to only depend on observed values: Strong Ignorability and the Stable Unit Treatment Value Assumption (SUTVA), which are commonly employed in observational studies (Dehejia & Wahba, 2002; Sekhon, 2009; Stuart, 2010; Austin, 2011). Strong Ignorability asserts that there is a non-zero probability of each subject receiving treatment, and that—conditional on covariates—the treatment assignment is independent of the potential outcomes (Rosenbaum & Rubin, 1983). SUTVA asserts that the potential outcomes of any subject i depends on $\mathbf{W}$ only through W_i and not other subjects' assignment (Rubin, 1980). Researchers can conduct sensitivity analyses to assess if treatment effect estimates are sensitive to violations of Strong Ignorability (e.g., Rosenbaum 2002, Chapter 4). See Sobel (2006), Hudgens & Halloran (2008), and Tchetgen & VanderWeele (2012) for a review of methodologies that address SUTVA violations.

2.2 Formalizing As-If Randomized Assignment in Matched Datasets

Strong Ignorability implies that $P(\mathbf{W}|\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X}) = P(\mathbf{W}|\mathbf{X})$, where $P(\cdot)$ denotes the probability measure on the set of all possible treatment assignments $\mathbb{W} \equiv \{0, 1\}^N$ (Imbens & Rubin, 2015). Thus, to conduct causal analyses assuming Strong Ignorability, researchers must assume $\mathbf{W} \sim P(\mathbf{W}|\mathbf{X})$ for some assignment mechanism $P(\mathbf{W}|\mathbf{X})$, i.e., that subjects in the matched dataset are as-if randomized according to a particular assignment mechanism. For example, assuming a matched dataset approximates a completely randomized

1. Such notation implicitly assumes the Stable Unit Treatment Value Assumption (Rubin, 1980), which we discuss shortly.

experiment equates to assuming, for a number of treated subjects N_T,

$$\textbf{Complete Randomization}: P(\mathbf{W}=\mathbf{w}|\mathbf{X}) = \begin{cases} \binom{N}{N_T}^{-1} & \text{if } \sum_{i=1}^{N} w_i = N_T \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

As another example, assuming a matched dataset approximates a blocked or paired randomized experiment equates to assuming that, for blocks (or pairs) $\mathcal{B}_1, \ldots, \mathcal{B}_J$,

$$\textbf{Block Randomization}: P(\mathbf{W}=\mathbf{w}|\mathbf{X}) = \begin{cases} \left[\prod_{j=1}^{J} \binom{N_j}{N_{jT}}\right]^{-1} \text{ if } \sum_{i\in\mathcal{B}_j} w_i = N_{jT} \ \forall j = 1, \ldots, J \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

where $N_j \equiv |\mathcal{B}_j|$ and N_{jT} denotes the number of treated subjects in $\mathcal{B}_j$.

Complete Randomization (1) and Block Randomization (2) are commonly assumed when conducting causal inference for matched datasets. For example, (2) is commonly assumed after using pairwise or blockwise matching algorithms, such as $1:k$ nearest neighbor matching (Rubin, 1973), $1:k$ optimal matching (Rosenbaum, 1989; Ming & Rosenbaum, 2001; Lu et al., 2011), full matching (Rosenbaum, 1991; Gu & Rosenbaum, 1993; Hansen, 2004; Hansen & Klopfer, 2006), and coarsened exact matching (Iacus et al., 2011, 2012). However, some argue that pairwise or blockwise matching is only a conduit to obtain group-level covariate balance, and thus (1) can be assumed instead (Ho et al., 2007; Schafer & Kang, 2008; Stuart, 2010). The assignment mechanisms (1) and (2) are also commonly assumed after using matching algorithms that optimize for group-level covariate balance (Zubizarreta, 2012; Zubizarreta et al., 2014; Kilcioglu & Zubizarreta, 2016; Zubizarreta & Keele, 2017).

As discussed in Section 1, researchers typically justify assuming Complete Randomization or Block Randomization by demonstrating that the matched dataset exhibits covariate balance, e.g., standardized covariate mean differences are below 0.1. The standardized covariate mean difference of the kth covariate is defined as (Austin, 2009a):

$$\bar{x}^*_{T,k} - \bar{x}^*_{C,k} = \frac{\bar{x}_{T,k} - \bar{x}_{C,k}}{\sqrt{\frac{s^2_{T,k}+s^2_{C,k}}{2}}}, \qquad k = 1, \ldots, K \tag{3}$$

where $\bar{x}_{T,k}$ and $\bar{x}_{C,k}$ are the sample means of the kth covariate in the treatment and control group, respectively, and $s^2_{T,k}$ and $s^2_{C,k}$ are the sample variances of the kth covariate in the treatment and control group, respectively. Thus, the standardized covariate mean differences are defined as the covariate mean difference divided by the pooled standard deviation (Flury & Riedwyl, 1986).[2] We use $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ to denote the K-length vector of

2. Following previous recommendations (Rosenbaum & Rubin, 1985; Stuart, 2008, 2010), we define the pooled standard deviation using the full, unmatched dataset (i.e., the dataset before matching). Thus, the use of $s^2_{T,k}$ and $s^2_{C,k}$ in (3) is a slight abuse of notation, because these sample variances are defined using the full, unmatched dataset, while $\bar{x}_{T,k}$ and $\bar{x}_{C,k}$ are defined using the matched dataset. As a result, the standardized covariate mean differences across different datasets will have the same denominator. Thus, a smaller standardized covariate mean difference denotes an improvement in mean balance, rather than a change in variance or sample size (Austin, 2014). In practice we have found that using the pooled standard deviation within the matched dataset (rather than the full, unmatched dataset) often gives highly similar results, which may explain the inconsistency across `R` packages: The popular packages `CBPS` (Imai & Ratkovic, 2014), `cobalt` (Greifer, 2017), `designmatch` (Zubizarreta & Kilcioglu, 2016),

covariate mean differences and $(\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)$ to denote the K-length vector of standardized covariate mean differences.

The most common way to assess covariate balance is to create a Love plot (Ahmed et al., 2006; Austin, 2009a; Zubizarreta, 2012), boxplot (Hansen, 2004; Rosenbaum, 2012), or table (Dehejia & Wahba, 1999; Harder et al., 2010) of the standardized covariate mean differences $(\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)$, but other diagnostics like significance tests (Smith & Todd, 2005; Lee, 2013), graphical displays of propensity score overlap (Rubin & Thomas, 2000; Dehejia & Wahba, 2002; Crump et al., 2009; Li et al., 2018a), and machine learning metrics (Linden & Yarnold, 2016; Gagnon-Bartsch & Shem-Tov, 2019) are also common. Thus, researchers will often iteratively match subjects until covariate balance diagnostics are satisfactory, but there has also been a recent surge in algorithms that ensure balance diagnostics are satisfactory by design in one step instead of iterative steps (Iacus et al., 2012; Zubizarreta, 2012; Imai & Ratkovic, 2014; Vegetabile et al., 2019). However, neither Complete Randomization nor Block Randomization fully conditions on the balance criteria that modern matching algorithms are designed to provide, whether it be iteratively or in one step. For example, the following assignment mechanism (which we call *Constrained Randomization*) conditions on the standardized covariate mean differences being below 0.1:

$$\textbf{Constrained Randomization} : P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \begin{cases} |\mathcal{A}|^{-1} & \text{if } \mathbf{w} \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $\mathcal{A} \equiv \{\mathbf{w} : |\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*| < 0.1 \text{ and } \sum_{i=1}^N w_i = N_T\}$ denotes the set of *constrained randomizations*. Constrained Randomization is similar to rerandomization, where subjects are randomized until covariate balance is achieved (Morgan & Rubin, 2012; Li et al., 2018b; Branson & Shao, 2021). To our knowledge, Constrained Randomization has not been used in the design and analysis of matched data, and we will explore its use throughout this paper.

How do we know when it is appropriate to assume Complete Randomization, Block Randomization, Constrained Randomization, or some other assignment mechanism for a matched dataset? While there are many covariate balance diagnostics and many matching algorithms designed to satisfy those diagnostics, current diagnostics do not assess which assignment mechanism should be assumed for a matched dataset. In Section 2.3, we present a randomization test for the hypothesis that treatment within a matched dataset follows a particular assignment mechanism (e.g., Complete Randomization, Block Randomization, or Constrained Randomization). In addition to appraising the covariate balance of a matched dataset, our randomization test procedure allows researchers to determine which assignment mechanism—if any—is most appropriate for a matched dataset, as we show in Section 2.4.

2.3 A Test for As-If Randomized Assignment in Matched Data

Here we outline a randomization test for the hypothesis that $H_0 : \mathbf{W} \sim P^*(\mathbf{W}|\mathbf{X})$ in a matched dataset. We use the notation $P^*(\mathbf{W}|\mathbf{X})$ to denote that this is a distribution

and `MatchIt` (Ho et al., 2011) use the pooled standard deviation within the full, unmatched dataset, while `Matching` (Sekhon, 2008) and `twang` (Ridgeway et al., 2017) use the pooled standard deviation within the matched dataset. Although we use the former throughout this paper, our `randChecks R` package allows researchers to use either pooled standard deviation of their choice.

posited by the researcher, rather than the *true* distribution of treatment assignment, which is never known in an observational study. The intuition behind our test is that it computes the distribution of balance we would expect if we conducted a randomized experiment on the data at hand using $P^*(\mathbf{W}|\mathbf{X})$ as the assignment mechanism. If the observed balance is substantially within the distribution of balance we would expect from a particular experimental design, then we do not find evidence against assuming that design. The test is as follows.

α-level Randomization Test for $H_0 : \mathbf{W} \sim P^*(\mathbf{W}|\mathbf{X})$ in a Matched Dataset

1. Specify an assignment mechanism $P^*(\mathbf{W}|\mathbf{X})$, which defines H_0.
2. Define a test statistic $t(\mathbf{W}, \mathbf{X})$, which measures covariate balance.
3. Generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)} \sim P^*(\mathbf{W}|\mathbf{X})$, where M is reasonably large (e.g., 1,000) to approximate the randomization distribution.
4. Compute the randomization distribution of covariate balance:
$$\left(t(\mathbf{w}^{(1)}, \mathbf{X}), \ldots, t(\mathbf{w}^{(M)}, \mathbf{X})\right) \tag{5}$$
5. Compute the following randomization-based two-sided p-value:
$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{I}(|t(\mathbf{w}^{(m)}, \mathbf{X})| \geq |t^{obs}|)}{M+1}, \quad \text{where } t^{obs} \equiv t\left(\mathbf{W}^{obs}, \mathbf{X}\right) \tag{6}$$
6. Reject H_0 if $p \leq \alpha$.

An immediate benefit of the above randomization test is that it is a valid and exact test for the null hypothesis that a specific assignment mechanism holds within a population of matched subjects; this readily follows from classical results on the validity of randomization tests (Edgington & Onghena, 2007; Good, 2013; Hennessy et al., 2016; Branson & Bind, 2019). Thus, our randomization test can be used to validly assess the plausibility of a given experimental design: If one rejects the null hypothesis H_0 for an assignment mechanism $P^*(\mathbf{W}|\mathbf{X})$, then that mechanism is not appropraite for the matched dataset. As discussed in Section 1, failing to reject the null hypothesis does not "prove" that random assignment holds, but it nonetheless serves as evidence that assuming a particular assignment mechanism for a matched dataset may be appropriate. Within the test, we recommend setting $\alpha = 0.15$, following common recommendations of other balance tests (Cattaneo et al., 2015; Hartman & Hidalgo, 2018); we elaborate on this point at the end of this section.

Our randomization test requires the researcher to specify a test statistic $t(\mathbf{W}, \mathbf{X})$ that measures covariate balance. Importantly, $t(\mathbf{W}, \mathbf{X})$ is not a function of the outcomes, which prevents researchers from biasing results when designing the matched dataset (Rubin, 2007, 2008b). The power of our randomization test depends on the relevance of the test statistic; see Rosenbaum (2002) and Imbens & Rubin (2015) for discussions of test statistic choices for randomization tests. We will focus on the standardized covariate mean differences (Stuart, 2010; Zubizarreta, 2012) and Mahalanobis distance (Mahalanobis, 1936; Rosen-

baum & Rubin, 1985; Gu & Rosenbaum, 1993; Diamond & Sekhon, 2013), because they are the most commonly used measures for appraising balance in matched datasets. Using the standardized covariate mean differences as a test statistic—i.e., defining K test statistics $t_1(\mathbf{W}, \mathbf{X}), \ldots, t_K(\mathbf{W}, \mathbf{X})$, where $t_k(\mathbf{W}, \mathbf{X}) = (\bar{x}^*_{T,k} - \bar{x}^*_{C,k})$ defined in (3)—results in a randomization test p-value for each covariate, thereby allowing for covariate-by-covariate assessments. Alternatively, one could set $t(\mathbf{W}, \mathbf{X}) = \max |(\bar{\mathbf{x}}^*_T - \bar{\mathbf{x}}^*_C)|$ (or another function of the standardized covariate mean differences) to define a single test statistic and p-value across covariates. As a global assessment of covariate balance, we will focus on setting $t(\mathbf{W}, \mathbf{X})$ equal to the Mahalanobis distance, which is defined as[3]

$$M = (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^T \left[\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\right]^{-1} (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C), \tag{7}$$

Note that the Mahalanobis distance is defined using the covariate mean differences $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$, rather than the standardized covariate mean differences $(\bar{\mathbf{x}}^*_T - \bar{\mathbf{x}}^*_C)$. In Section 2.4 we demonstrate how to create easy-to-interpret graphical displays of our test using the standardized covariate mean differences or the Mahalanobis distance. In general we recommend using the Mahalanobis distance as a test statistic, because it accounts for the joint behavior of covariates, instead of only assessing marginal balance for each covariate.

Our randomization test also requires generating random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)} \sim P^*(\mathbf{W}|\mathbf{X})$. Sometimes these draws can be generated via permutations of the observed treatment assignment $\mathbf{W}^{obs}$; this is the case for generating draws from Complete Randomization in (1) and Block Randomization in (2). In other cases, these draws can be generated via rejection-sampling: For example, to generate draws from Constrained Randomization in (4), one can generate draws from Complete Randomization in (1) and only accept a draw if $|\bar{\mathbf{x}}^*_T - \bar{\mathbf{x}}^*_C| < 0.1$. If rejection-sampling is computationally intensive, importance-sampling can be used to approximate the randomization test p-value (Branson & Bind, 2019).

Our test is a generalization of other balance tests. For example, Hansen & Bowers (2008) and Hansen (2008) proposed permutation tests using the Mahalanobis distance as a test statistic, and Gagnon-Bartsch & Shem-Tov (2019) proposed a permutation test using machine-learning methods to construct a test statistic. Another test is the Cross-Match test (Rosenbaum, 2005), which focuses on the pairwise nature of matched datasets. Permutation tests have also been used to assess the balance of subjects in regression discontinuity designs (Cattaneo et al., 2015; Mattei & Mealli, 2016). All of these tests are special cases of our randomization test, where draws from $P^*(\mathbf{W}|\mathbf{X})$ correspond to permutations of $\mathbf{W}^{obs}$.

However, as noted elsewhere in the literature (e.g., Cattaneo et al. (2015) and Hartman & Hidalgo (2018)), Type II errors are a concern for balance tests like ours, because we want to avoid falsely concluding that treatment is effectively randomized when really it is not. One option for avoiding Type II errors is to set α to be larger than 0.05 (e.g., Cattaneo et al. (2015) recommend setting $\alpha = 0.15$, and Hartman & Hidalgo (2018) noted that many researchers choose 0.15 or 0.2). Hartman & Hidalgo (2018) recommend using equivalence tests instead of balance tests, which essentially "flip the null and alternative," such that

3. Here, $\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) = \frac{N}{N_T N_C}\text{cov}(\mathbf{X})$, where $\text{cov}(\mathbf{X})$ is the sample covariance matrix of $\mathbf{X}$. This equality is derived in Morgan & Rubin (2012). Furthermore, similar to our definition of the standardized covariate mean differences in (3)—where we recommend using the pooled standard deviation within the full, unmatched dataset—we use the full, unmatched dataset when computing $\text{cov}(\mathbf{X})$, but our `R` package `randChecks` also allows researchers to define $\text{cov}(\mathbf{X})$ using the matched dataset if they desire.

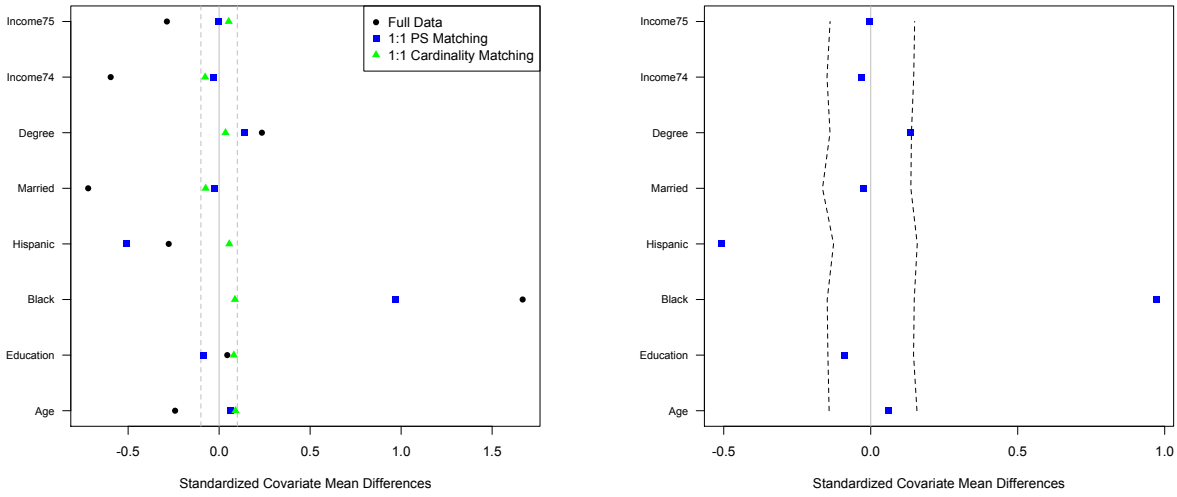
the null hypothesis is that the data are not randomized and the alternative is that they are randomized, thereby avoiding the issues of a test being underpowered. The equivalence tests of Hartman & Hidalgo (2018) are a promising way to assess randomized treatment assignment, but Hartman & Hidalgo (2018) focused only on covariate-by-covariate assessments of Complete Randomization, and it is unclear how to extend their approach to other assignment mechanisms or omnibus metrics like the Mahalanobis distance.

Relatedly, the power of our randomization test depends on—in addition to the test statistic—the alternative hypothesis that holds, if not the randomized assignment null hypothesis $H_0 : \mathbf{W} \sim P^*(\mathbf{W}|\mathbf{X})$. Thus, studying the power of balance tests like ours requires quantifying the discrepancy between random assignment and the true assignment mechanism. For example, Hansen (2009) quantified the discrepancy between propensity score matched data and block randomization using the within-block variability of propensity scores and covariates, and established that the power of balance tests like ours tends to one if this variability is asymptotically large (e.g., Proposition 2.1 of Hansen (2009)). An interesting avenue for future work is characterizing the discrepancy between other designs—such as Constrained Randomization—and alternative assignment mechanisms that may hold in matched data, thereby providing a way to study the power of balance tests like ours.

2.4 Graphical Diagnostics for Assessing Different Designs: The Lalonde Data

As an example of how to use our randomization test—and how graphical diagnostics can complement our test—we will consider the Lalonde dataset (LaLonde, 1986), which has been extensively used for evaluating matching methods (Dehejia & Wahba, 1999, 2002; Smith & Todd, 2005; Iacus et al., 2012; Diamond & Sekhon, 2013). The data are available in the `MatchIt R` package (Stuart et al., 2011); the treatment group consists of 185 individuals who participated in the National Supported Work Demonstration, and the control group consists of 429 individuals sampled from the Population Survey of Income Dynamics. There are eight covariates: age, years of education; whether someone is black, whether someone is Hispanic, whether or not someone is married, whether or not someone has a high school degree, and income in 1974 and 1975. We defer to the aforementioned references for fuller descriptions of the data, since they have been extensively studied previously. The tests and visuals in this section are reproduced as examples in our `randChecks R` package; all the code used to conduct these tests and create these visuals is provided in Appendix A.

We will assess the balance of three datasets: The full Lalonde data, a dataset from 1:1 propensity score matching, and a dataset from 1:1 cardinality matching. We implemented 1:1 propensity score matching using the `MatchIt R` package and a logistic regression to estimate the propensity scores. Meanwhile, we implemented cardinality matching using the `designmatch R` package (Zubizarreta & Kilcioglu, 2016), which finds the largest subset of the observational data that fulfills prespecified covariate balance constraints (Zubizarreta et al., 2014). In our implementation of cardinality matching, we required the eight standardized covariate mean differences to be less than 0.1. In Sections 4 and 5 we focus on cardinality matching, because it guarantees that covariate balance constraints hold within a matched dataset. However, cardinality matching may discard treated subjects to achieve balance constraints: The 1:1 propensity score matched dataset contained 370 subjects, while the 1:1 cardinality matched dataset contained 240 subjects.



(a) Love plot for the full Lalonde dataset, the 1:1 propensity score matched dataset, and the 1:1 cardinality matching dataset.

(b) Love plot for the 1:1 propensity score matched dataset with 7.5% and 92.5% complete randomization quantiles denoted by dashed lines.

Figure 1: Assessing balance of matched datasets using Love plots.

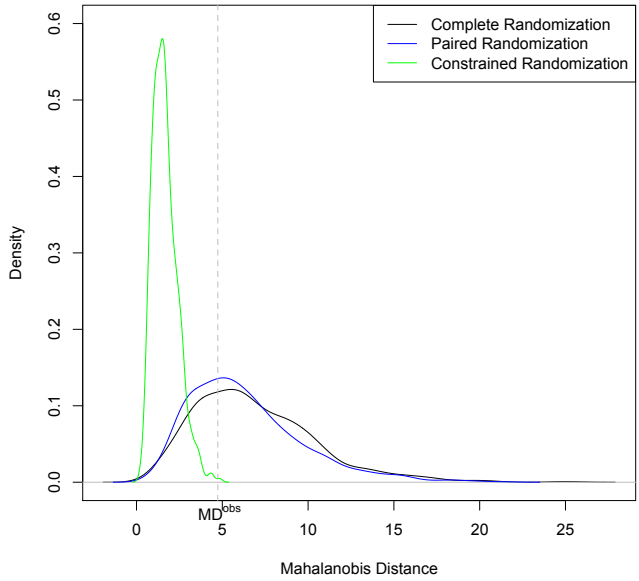


Figure 2: Distribution of the Mahalanobis distance across 1,000 complete randomizations, paired randomizations, and constrained randomizations for the 1:1 cardinality matched dataset. The observed Mahalanobis distance is denoted by a dashed vertical line.

Figure 1a shows the standardized covariate mean differences for these three datasets. There are large imbalances in the full dataset, which motivates matching methods. The matched datasets both exhibit degrees of covariate balance. First we ask: Is it reasonable to assume the propensity score matched dataset approximates a completely randomized experiment? To assess this, we run our randomization test under Complete Randomization by permuting $\mathbf{W}^{obs}$ 1,000 times and computing the standardized covariate mean differences for each permutation. Figure 1b shows the 7.5% and 92.5% quantiles of each difference across these permutations; this corresponds to setting $\alpha = 0.15$ within our randomization test, using the standardized covariate mean differences as the test statistic. Any difference outside of these quantiles is considered surprisingly large under Complete Randomization. Because there are some differences outside of these quantiles, we conclude that Complete Randomization is not plausible for this dataset. Running our test using the Mahalanobis distance as a test statistic gave the same conclusion: The p-value was less than 0.001.

Now we ask: Is it reasonable to assume the cardinality matched dataset approximates a completely randomized experiment, or even another experimental design? Figure 2 provides a visual of our test using the Mahalanobis distance as a test statistic. We display the distribution of the Mahalanobis distance across 1,000 draws from Complete Randomization in (1), 1,000 draws from Paired Randomization in (2), and 1,000 draws from Constrained Randomization in (4). We see that the observed Mahalanobis distance is reasonably within the first two distributions, giving credence to assuming Complete Randomization or Paired Randomization (the corresponding p-values are 0.694 and 0.636, respectively). However, the observed Mahalanobis distance is almost completely outside of the Constrained Randomization distribution, suggesting that this experimental design is not plausible (the corresponding p-value is 0.002). At first this may be surprising: Constrained Randomization in (4) incorporates the constraint that all standardized covariate mean differences are below 0.1, and cardinality matching fulfills this constraint, so why does this design not appear plausible? There are two reasons. First, even though cardinality matching fulfilled this balance constraint, it did so just barely, as seen in Figure 1a. Thus, the matched dataset is unusually imbalanced according to this design. Second, the Mahalanobis distance accounts for the covariance among covariates; thus, the *joint* imbalance in Figure 1a is considered unusual according to this design.

The above example demonstrates how our randomization test can make covariate-by-covariate or omnibus assessments of balance. In particular, we recommend using an omnibus assessment via the Mahalanobis distance, because it accounts for the joint relationship among covariates when assessing balance. Furthermore, using the Mahalanobis distance allows for a graphical display that places different designs on the same univariate scale, such that researchers can ascertain which design is most appropriate for a particular dataset. Finally, this example demonstrates that even when covariate balance constraints hold for a matched dataset, it still may not be appropriate to assume a balance-constrained design.

3. The Analysis Stage: After Assuming an Experimental Design for Matched Data

Once a particular design is assumed for a matched dataset, causal analyses become relatively straightforward, to the extent that they are straightforward for an experiment that uses

that design. There are Fisherian (also known as randomization-based), Neymanian, and Bayesian modes of inference for analyzing such experiments.

Randomization-based inference focuses on testing sharp hypotheses, such as $Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$. Under a sharp hypothesis, the potential outcomes for any treatment assignment are known. Researchers can also invert sharp hypotheses to obtain point estimates and uncertainty intervals. One possible sharp hypothesis is that the treatment effect is additive, i.e., that $Y_i(1) = Y_i(0) + \tau$ for some $\tau \in \mathbb{R}$ for all $i = 1, \ldots, N$. Then, a randomization-based uncertainty interval is the set of τ such that one fails to reject this sharp hypothesis, and a randomization-based point estimate is the τ yielding the highest p-value (Hodges Jr & Lehmann, 1963; Rosenbaum, 2002). To test such hypotheses, one must specify an assignment mechanism (and methods from Section 2 can be used to specify this) and a test statistic (some kind of treatment effect estimator). The choice of assignment mechanism can be viewed as a design-based decision, and the choice of test statistic can be viewed as a model-based decision. See Rosenbaum (2002) and Imbens & Rubin (2015) for a general discussion of randomization-based inference. A limitation of this mode of inference is that a sharp null hypothesis must be specified, which often requires assuming away heterogeneous treatment effects, as in the aforementioned additive sharp hypothesis. See Samii & Aronow (2012, Section 4) for further discussion on this limitation and Caughey et al. (2016) for an approach that incorporates treatment effect heterogeneity for this mode of inference. Despite this limitation, because randomization-based inference can flexibly incorporate different design-based and model-based decisions, we will focus primarily on this mode of inference in Sections 4 and 5.

Meanwhile, the Neymanian mode of inference involves asymptotic approximations for treatment effect estimators under certain experimental designs. There are well-known results on Neymanian inference for many experimental designs. For example, Miratrix et al. (2013) discusses Neymanian inference for blocked experiments, and Imai (2008) does the same for paired experiments. See Pashley & Miratrix (2017) for a discussion of variance estimation for these two designs as well as hybrid designs that involve blocks and pairs. Neymanian inference has also been established for factorial designs (Dasgupta et al., 2015), rerandomized experiments (Li et al., 2018b), and their combination (Li et al., 2020). See Ding (2017) for a comparison of randomization-based and Neymanian modes of inference for blocked, matched-pair, and factorial designs, as well as Fogarty (2020, Section 1) for a historical, agnostic view on these modes of inference.

The seminal paper by Li et al. (2018b) is particularly relevant to this work. Li et al. (2018b) derived the asymptotic distribution of the mean-difference estimator for rerandomized experiments where the Mahalanobis distance is constrained to be below some threshold (an experimental design first proposed by Morgan & Rubin (2012)). This rerandomization scheme is similar to (but not the same as) Constrained Randomization (defined in (4)). A promising line for future work is extending the results of Li et al. (2018b) to designs like Constrained Randomization. For example, Wang & Zubizarreta (2019) established several large-sample properties of matching methods that ensure covariate balance constraints hold by design. Continuing such developments would provide a useful way to conduct Neymanian inference for matched datasets using nuanced designs like Constrained Randomization.

Finally, the Bayesian mode of inference for estimating causal effects was first formalized in Rubin (1978). Under this mode of inference, $(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ are treated as unknown,

and models for these quantities must be specified. After assuming $\mathbf{W} \sim P^*(\mathbf{W}|\mathbf{X})$, the remaining work for a Bayesian analysis is to specify statistical models for the covariates and outcomes. This mode of inference is particularly useful for incorporating uncertainty in many complex data scenarios, such as noncompliance (Frangakis et al., 2002), missing data (Rubin, 1996), their combination (Barnard et al., 2002), and multi-site trials (Dehejia, 2003); however, these complications are outside the scope of this work. See Imbens (2004) and Heckman et al. (2014) for discussions of Bayesian inference for randomized experiments.

In what follows, we demonstrate how our randomization test can assess the plausibility of experimental designs for matched datasets, first in simulation (Section 4) and then in a real data analysis (Section 5). In these sections we also consider standard Neymanian approaches for analyzing matched datasets assuming complete randomization or paired randomization, as well as a randomization-based approach for analyzing matched datasets assuming designs that leverage covariate balance, such as constrained randomization.

4. Simulations

In this section, we will simulate datasets that exhibit covariate imbalance, thereby motivating matching methods. We will demonstrate how our test in Section 2 can be used to assess complete randomization, paired randomization, and constrained randomization designs for matched datasets. Furthermore, we will demonstrate how assuming constrained randomization for matched datasets can improve the precision of causal analyses when researchers match on covariates that are related to the outcomes. However, we will also see that assuming a precise experimental design can harm inferential results if there are still substantial biases due to covariate imbalances that remain after matching.

4.1 Simulation Setup

We follow the simulation setup of Austin (2009b) and Resa & Zubizarreta (2016), which has been used to evaluate different matching methods. Consider a dataset with $N_T = 250$ treated subjects and $N_C = 500$ control subjects. Each subject has four Normally distributed covariates and four Bernoulli distributed covariates. These eight covariates are generated such that the true standardized mean difference is 0.2 for half of the covariates and 0.5 for the other half. Furthermore, there are three outcomes: The first outcome is a linear function of the covariates, and the other two are nonlinear functions, where the third outcome is a more complex function than the second.[4] For each outcome, there is an additive treatment effect of one, which is the causal estimand in this simulation. Details about this data generating process are in Appendix B. By repeating this data-generating process, we produced 1,000 datasets with severe covariate imbalances.

4.2 Design and Analysis Results for Complete Randomization and Paired Randomization

In their simulation study, Resa & Zubizarreta (2016) compared nearest-neighbor propensity score matching, optimal subset matching (Rosenbaum, 2012), and cardinality matching

4. When each outcome was regressed on the covariates, $R^2 = 0.9$, $R^2 = 0.5$, and $R^2 = 0.25$ for the first, second, and third outcomes, respectively, on average across the 1,000 replicated datasets.

(Zubizarreta et al., 2014). Cardinality matching is similar to optimal subset matching in that it may discard some treated subjects in the name of achieving better balance; however, it differs from optimal subset matching in that it ensures group-level balance directly. Resa & Zubizarreta (2016) found that cardinality matching performs better than nearest-neighbor and optimal subset matching in terms of covariate balance, sample size, bias, and root mean squared error (RMSE). Thus, we focus on cardinality matching, and defer to Resa & Zubizarreta (2016) and other simulation studies (e.g., Austin (2009b) and Austin (2014)) for a comparison of other methods.

When implementing cardinality matching, we focus on creating the largest pair-matched dataset such that the absolute standardized covariate mean differences are less than some threshold a; we consider $a = 0.1$ and $a = 0.01$. A more stringent threshold results in better balance but possibly a smaller sample size.[5] Doing this produced 1,000 matched datasets with threshold $a = 0.1$ and 1,000 datasets with threshold $a = 0.01$. Table 1 shows the bias, variance, and RMSE of the mean-difference estimator across these matched datasets. Table 1 also shows the coverage of what we call "complete randomization 95% confidence intervals" and "paired randomization 95% confidence intervals," which are computed as

$$\underbrace{\hat{\tau} \pm 1.96\sqrt{\frac{\hat{\sigma}_T^2}{N_T} + \frac{\hat{\sigma}_C^2}{N_C}}}_{\text{complete randomization}} \qquad \text{and} \qquad \underbrace{\hat{\tau} \pm 1.96\sqrt{\frac{\sum_{j=1}^{J}(\hat{\tau}_j - \hat{\tau})^2}{J(J-1)}}}_{\text{paired randomization}} \tag{8}$$

where $\hat{\tau}$ is the mean-difference estimator, $\hat{\tau}_j$ is the mean-difference estimator within pair $j = 1, \ldots, J$, and $\hat{\sigma}_T^2$ and $\hat{\sigma}_C^2$ are the sample variances in treatment and control. These are the standard Neymanian confidence intervals for the average treatment effect in completely randomized and paired experiments (Imai, 2008; Imbens & Rubin, 2015).

When $a = 0.1$, there is substantial bias, resulting in confidence intervals undercovering. As expected, the paired randomization confidence intervals are narrower than the complete randomization confidence intervals, resulting in even worse coverage. Thus, it can be harmful to condition on the pairs of a matched dataset. However, when $a = 0.01$, the bias is neglible and the confidence intervals tend to overcover. This overcoverage is most prominent for the first outcome, followed by the second, and finally by the third. This is ordered by how "well-specified" cardinality matching was, which only attempted to balance the raw covariates—which define the first outcome—and not the nonlinear functions that define the second and third outcomes (as detailed in Appendix B).

Table 1 show the inferential results when we assume complete randomization or paired randomization for the matched datasets, but were these assumptions appropriate? Note that all of the matched datasets fulfill the common rule-of-thumb that standardized covariate mean differences be below 0.1, and thus it would be common practice to assume complete randomization or paired randomization for these datasets. Alternatively, it is also common to perform balance tests to assess these assumptions; we will consider three here:

5. When $a = 0.1$, sample sizes ranged from 438 to 500, and when $a = 0.01$, they ranged from 386 to 494. A sample size less than 500 means some treated subjects were discarded in the name of achieving better balance. Discarding treated subjects changes the causal estimand, but this isn't problematic when the treatment effect is homogeneous (as is the case here).

1. Use a t-test for each of the eight covariates, and then define the balance test p-value as the minimum of the eight resulting p-values.

2. Use a Kolmogorov-Smirnov (KS) test for each of the eight covariates, and then define the balance test p-value as the minimum of the eight resulting p-values.

3. Use our randomization test from Section 2.3, using the KS statistic as the test statistic for each of the eight covariates, and then define the balance test p-value as the minimum of the eight resulting p-values.

4. Use our randomization test from Section 2.3, using the Mahalanobis distance (7) as the test statistic.

For the first three balance tests, we choose the minimum p-value among the covariate-specific p-values in order to be conservative when assessing randomized assignment (Diamond & Sekhon, 2013; Cattaneo et al., 2015). Meanwhile, our randomization test using the Mahalanobis distance acts as a global test for covariate balance, thereby providing a single p-value. Table 2 shows the rejection rate of complete randomization and paired randomization for these balance tests, where we reject if the p-value is less than $\alpha = 0.15$ (as recommended in Cattaneo et al. (2015) and our Section 2.3).

For the $a = 0.1$ matched datasets, the t-test never rejects complete randomization or paired randomization, where we used the paired t-test to assess paired randomization. Meanwhile, the KS test and our test reject complete randomization about half the time. Interestingly, our test rejects complete randomization most frequently when we use the KS statistic—thus, it may be beneficial to use the exact distribution of the KS statistic (as implemented in our test) rather than the asymptotic distribution (as implemented in the KS test). Furthermore, note that, unlike the t-test, there is not a paired version of the KS test, and thus we do not display a paired randomization p-value. On the other hand, our test can assess any experimental design, and thus we can still use the KS statistic to test paired randomization. We see that our test frequently rejects paired randomization, where using the Mahalanobis distance resulted in slightly more power. Thus, within our test, the KS statistic was more powerful for assessing complete randomization, while the Mahalanobis distance was more powerful for assessing paired randomization. An interesting direction for future research is exploring test statistics that maximize power for detecting violations of different experimental designs. In any case, because inferential results are quite biased when $a = 0.1$ (as shown in Table 1), it is reassuring that our test and the KS test frequently reject randomized assignment for these matched datasets. This demonstrates the advantages of assessing forms of covariate balance beyond marginal means—e.g., joint mean balance like the Mahalanobis distance or marginal distribution balance via the KS statistic. At the same time, this demonstrates that our approach is not a panacea: Another way to view Table 2 is that our test incorrectly failed to reject complete randomization nearly half the time and paired randomization nearly a quarter of the time, thereby giving us false confidence in biased analyses. Thus, this also demonstrates that even if our test fails to reject a particular design, there's no guarantee that the resulting inference using such a design will be valid or unbiased.

Meanwhile, for the $a = 0.01$ matched datsaets, the t-test and our test using the Mahalanobis distance never reject complete randomization or paired randomization, although

Outcome	Bias	Variance	RMSE	Coverage (CR 95% CIs)	Coverage (PR 95% CIs)
First Outcome					
$a = 0.1$	0.92	0.04	0.94	73.5%	51.3%
$a = 0.01$	0.05	0.04	0.20	100%	100%
Second Outcome					
$a = 0.1$	2.10	0.58	2.24	58.1%	50.1%
$a = 0.01$	0.42	0.54	0.84	99.7%	99.1%
Third Outcome					
$a = 0.1$	0.79	1.56	1.48	92.1%	90.5%
$a = 0.01$	0.01	1.40	1.18	98.1%	97.5%

Table 1: Properties of the mean-difference estimator after cardinality matching, where the standardized covariate mean differences are constrained to be less than some threshold a. Complete randomization (CR) and paired randomization (PR) intervals are defined in (8).

Dataset/Test	CR p-value Rejection Rate	PR p-value Rejection Rate
$a = 0.1$		
t-test	0.0%	0.0%
Kolmogorov-Smirnov test	48.7%	NA
Randomization test (KS)	60.8%	65.5%
Randomization test (MD)	45.2%	76.7%
$a = 0.01$		
t-test	0.0%	0.0%
Kolmogorov-Smirnov test	3.2%	NA
Randomization test (KS)	5.5%	8.5%
Randomization test (MD)	0.0%	0.0%

Table 2: Rejection rate for four balance tests—where we reject if the p-value is less than $\alpha = 0.15$—for the $a = 0.1$ and $a = 0.01$ cardinality matched datasets. By using a two-sample or paired t-test, we can assess complete randomization (CR) or paired randomization (PR); however, because there is not a paired version of the Kolmogorov-Smirnov (KS) test, it can only assess CR. However, we can use the KS statistic as a test statistic within our randomization test, thereby allowing us to test CR or PR using the KS statistic. For the first three tests, to be conservative, we took the minimum of the eight covariate-specific p-values. Meanwhile, our test using the Mahalanobis distance (MD) assesses global covariate balance with a single p-value.

the KS test rejects complete randomization 3.2% of the time. Furthermore, our test using the Mahalanobis distance rejects complete randomization 5.5% of the time and paired randomization 8.5% of the time. Because inferential results are relatively unbiased and conservative for the $a = 0.01$ matched datasets, this lack of rejection is reassuring.

4.3 Design and Analysis Results for Constrained Paired Randomization

In the previous section we found minimal evidence against complete randomization or paired randomization for the $a = 0.01$ cardinality matched datasets, which were designed to exhibit a high level of balance across all covariates. Now we'll consider the consequences of analyzing these matched datasets assuming a variant of Constrained Randomization:

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \begin{cases} |\mathcal{A}|^{-1} & \text{if } \mathbf{w} \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where $\mathcal{A} \equiv \{\mathbf{w} : |\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*| < 0.05 \text{ and } \sum_{i=1}^N w_i = N_T\}$. This is a natural design to posit for cardinality matching, because we constrained $|\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*|$ by design. To assess if this design is plausible for the $a = 0.01$ cardinality matched datasets, we ran our test using the Mahalanobis distance as a test statistic; the randomization test p-values were consistently greater than 0.8, suggesting the plausibility of this experimental design.

To analyze the $a = 0.01$ cardinality matched datasets assuming constrained randomization in (9), we take a randomization-based approach and test the sharp hypothesis H_0^τ : $Y_i(1) = Y_i(0) + \tau$ for $\tau \in \{-12.00, -11.99, \ldots, 11.99, 12.00\}$. Define the 95% randomization-based confidence interval as the set of τ such that we fail to reject H_0^τ at the $\alpha = 0.05$ level. We'll consider the mean-difference estimator and linear regression estimator as test statistics when inverting this hypothesis test.

The first three rows of Table 3 compare the randomization-based confidence intervals for the mean-difference estimator under constrained randomization to the complete randomization and paired randomization confidence intervals from the previous section. The coverage of the constrained randomization confidence intervals is closer to the nominal level, and they are substantially narrower: They are on average 45% the width for the first outcome, 66% the width for the second outcome, and 88% the width for the third outcome, as compared to the complete randomization confidence intervals. However, the constrained randomization confidence intervals undercover for the second outcome. This is likely because of nonneglible bias, as seen in Table 1. Thus, in some sense it was lucky that the complete randomization and paired randomization confidence intervals overcovered for the second outcome despite this bias. This echoes the observation made in Table 1, where paired randomization undercovered even more so than complete randomization for the $a = 0.1$ cardinality matched datasets. Thus, it may be harmful to assume a precise experimental design or even any design if there are substantial biases that remain after matching. Indeed, an argument can be made to conservatively assume a less precise design—e.g., complete randomization—in the hope that the resulting wider confidence intervals will be closer to the nominal level if there are biases that remain after matching.

Meanwhile, the last three rows of Table 3 show results for the treatment effect estimator from a linear regression of the outcomes on $\mathbf{X}$ and $\mathbf{W}$. The complete randomization confidence intervals were computed using the standard error of the regression coefficient for $\mathbf{W}$. The paired randomization confidence intervals were computed using the variance estimator of Fogarty (2018), who provides recent results on regression adjustment for paired experiments.[6] In this case, the results across experimental designs are nearly identical.

6. We used the variance estimator S_{R1}^2 in Fogarty (2018), which utilizes pairwise-differences of functions of the covariates among matched pairs. We specified this function as the raw covariates, $(\mathbf{x}_1, \ldots, \mathbf{x}_8)$,

CI Method	*First Outcome*		*Second Outcome*		*Third Outcome*	
	Average CI Length	**Coverage**	**Average CI Length**	**Coverage**	**Average CI Length**	**Coverage**
Mean-Difference Analysis						
Complete Rand.	2.14	100%	4.55	99.7%	5.09	98.1%
Paired Rand.	1.93	100%	4.21	99.1%	4.95	97.5%
Constrained Rand.	0.96	97.9%	2.94	91.5%	4.46	95.5%
Linear Regression Analysis						
Complete Rand.	0.74	95.3%	2.71	90.8%	4.49	95.9%
Paired Rand.	0.74	95.2%	2.65	88.9%	4.46	95.6%
Constrained Rand.	0.73	94.3%	2.71	90.2%	4.40	95.4%

Table 3: Average length and coverage of complete randomization, paired randomization, and constrained randomization confidence intervals for the mean-difference estimator and linear regression estimator for the $a = 0.01$ cardinality matched datasets.

When the covariates are linearly related with the outcome—as is the case for the first outcome—coverage for linear regression is close to the nominal level. However, there is undercoverage for the second outcome, which is a nonlinear function of the covariates. This provides two findings. First, linear regression after matching is not guaranteed to exhibit the correct coverage. Second, as discussed in Section 3, the assignment mechanism in our randomization-based approach can be viewed as a design-based choice, while the treatment effect estimator is a model-based choice. Table 3 suggests that there may be an equivalence among certain design-based and model-based choices, which echoes recent equivalences found between rerandomization designs and regression adjustment (Li & Ding, 2020).

In summary, our test can assess standard experimental designs like complete randomization and paired randomization, as well as nuanced designs like constrained randomization. When matched datasets exhibit high levels of covariate balance, complete randomization and paired randomization analyses can be conservative, and designs that account for strong covariate balance can provide more precise causal analyses. However, using a precise experimental design like paired randomization or constrained randomization should be proceeded with caution, because it can harm inferential results if there are still substantial biases due to covariate imbalances that were not accounted for in the matching stage. Nonetheless, the additional precision from precise experimental designs can be substantial if researchers match on relevant covariates. In particular, subject-matter expertise often guides researchers towards balancing certain covariates that are deemed relevant a priori. In Section 5, we revisit a causal analysis conducted by Keele et al. (2017), who used matching to target balancing certain covariates. We will demonstrate how our randomization test can assess the type of experimental design their matched dataset approximates, and how a constrained randomization design can improve precision for this application.

to make the paired randomization analysis comparable to the complete randomization analysis. Fogarty (2018) also discusses utilizing pairwise-*averages* of functions of the covariates among matched pairs, which we do not consider here.

5. Revisiting a Causal Analysis of the Effects of Candidates' Race on Black Voter Turnout

An ongoing problem in political science is determining how minority candidates affect minority voter turnout in American elections. Many works have found a positive relationship between minority candidate participation in elections and minority voter turnout; to explain this phenomenon, these works argue that minorities feel empowered when they witness a minority candidate run for political office, thereby increasing voter turnout (Browning et al., 1984; Bobo & Gilliam, 1990; Leighley, 2001; Barreto et al., 2004). However, these findings have primarily been correlational instead of causal.

Recently, Keele et al. (2017) addressed this research question using matching to conduct a causal analysis assessing if having at least one African American candidate in Louisiana mayoral elections affected black voter turnout.[7] We will apply our randomization test to the Keele et al. (2017) matched dataset to assess if this dataset approximates a particular experimental design, use our randomization-based inferential approach to analyze the matched dataset, and compare our approach to the more standard approach used in Keele et al. (2017). Revisiting this causal analysis is particularly suitable for assessing our approach for two reasons. First, Keele et al. (2017) used cardinality matching, which is the method we focused on in Section 4. Second, and more importantly, Keele et al. (2017) used subject-matter expertise to target achieving high levels of balance on certain covariates, and our approach can condition on these high levels of covariate balance to provide a precise causal analysis. First, we describe the full data and matched data in Keele et al. (2017). Then, we compare our approach to a standard approach for analyzing these data.

5.1 Description of the Full Dataset and Matched Dataset

The data include 1,006 mayoral elections in Louisiana from 1988-2011. Data is at the municipality level, where each election was held. Covariates include each municipality's median income, number of residents, percentage of residents that are African American and of voting age, percentage of residents with a college degree, percentage of residents with a high school degree, percentage of residents that are unemployed, percentage of residents that are below the poverty line, and whether or not it had a home rule charter.[8] The treatment is whether or not at least one candidate in the election was African American. The outcome is black voter turnout (measured in percentage points), and interest is in the ATE on this outcome. Keele et al. (2017) created this dataset using three data sources maintained by the state of Louisiana, and further details can be found therein.

Keele et al. (2017) focused their efforts on these data because all mayoral elections in Louisiana can turn into a runoff election. In Louisiana mayoral elections, a "general election" is held at first, where any number of candidates may run. If no candidate receives the majority of votes, the two candidates with the most votes advance to a "runoff election." Keele et al. (2017) analyzed general elections as well as runoff elections. For ease of exposi-

7. Following the practice of the United States Census Bureau and Keele et al. (2017), we use "African American" and "black" interchangeably.
8. These covariates are based on 1990 census data and thus are pre-treatment measurements for most of the electoral data.

tion, we focus on general elections, because Keele et al. (2017) were able to achieve a larger matched sample and higher level of covariate balance for these data.

The full dataset exhibits large covariate imbalances (see Table 1 of Keele et al. (2017)). In particular, treated municipalities have substantially higher proportions of African American residents. Keele et al. (2017) posited that this covariate was particularly relevant to black voter turnout, and so they used cardinality matching to ensure strong balance on this covariate, as well as balance on the other covariates. Furthermore, they ensured near-exact balance (Rosenbaum, 2010, Chapter 9) on election year by creating pairs of treatment and control elections that occurred during the same year or one year after each other. This is an example of the common practice of using subject-matter expertise to prioritize certain forms of covariate balance when designing a matched dataset (Ramsahai et al., 2011; Zubizarreta, 2012; Pimentel et al., 2015; Keele & Small, 2020). The resulting matched dataset consisted of 197 pairs of elections that exhibited (1) near-exact balance on election year within each pair, (2) high balance on percentage of African American residents across pairs, and (3) balance on all other covariates across pairs. Table 4 shows the standardized covariate mean differences for each covariate in this dataset; all but one of the differences are below 0.1—the municipal population covariate had a standardized mean difference of 0.107.

5.2 Which Experimental Design Does This Matched Dataset Approximate?

Keele et al. (2017) argued that their matched dataset approximated a paired experiment by finding non-significant Kolmogorov-Smirnov tests for each covariate. This is a standard diagnostic in the matching literature, but it is not a valid test for a specific experimental design. To provide a valid test, we ran our randomization test for complete randomization and two paired designs:

$$P(\mathbf{W}=\mathbf{w}|\mathbf{X}) = \begin{cases} \binom{394}{197}^{-1} & \text{if } \sum_{i=1}^{394} w_i = 197 \\ 0 & \text{otherwise.} \end{cases} \text{(Complete Randomization)} \tag{10}$$

$$P(\mathbf{W}=\mathbf{w}|\mathbf{X}) = \begin{cases} \frac{1}{2^{197}} & \text{if } \sum_{i\in\mathcal{B}_j} w_i = 1 \ \forall j = 1,\ldots,197 \\ 0 & \text{otherwise.} \end{cases} \text{(Paired Randomization)} \tag{11}$$

$$P(\mathbf{W}=\mathbf{w}|\mathbf{X}) = \begin{cases} |\mathcal{A}^{(cp)}|^{-1} & \text{if } \mathbf{w}\in\mathcal{A}^{(cp)} \\ 0 & \text{otherwise.} \end{cases} \text{(Constrained Paired Randomization)} \tag{12}$$

where $\mathcal{A}^{(cp)} \equiv \{\mathbf{w} : \sum_{i\in\mathcal{B}_j} w_i = 1 \forall j = 1,\ldots,197, \ \ |\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*| < 0.15, \text{ and } |\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*|^{(AA)} < 0.01\}$ is the set of *constrained paired randomizations*, where $|\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*|^{(AA)}$ denotes the absolute standardized mean difference for the African American (%) covariate. Complete Randomization in (10) does not account for covariate balance; Paired Randomization in (11) recognizes that Keele et al. (2017) constructed pairs of elections with near-exact balance on election year; and Constrained Paired Randomization in (12) additionally recognizes that certain levels of balance were achieved across pairs, particularly for the African American (%) covariate. To run our randomization test, we generated 10,000 random draws from the three above assignment mechanisms and computed the standardized covariate mean differences for each draw; then, the randomization test p-value in (6) is defined as the proportion of absolute standardized covariate mean differences that are greater than the

Covariate	$\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$	CR p-value	PR p-value	Constr. PR p-value
Municipal Population	0.11	0.34	0.30	0.18
African American (%)	0.00	0.99	0.99	0.94
College degree (%)	0.00	0.96	0.94	0.93
High school degree (%)	0.00	0.96	0.94	0.93
Unemployed (%)	0.02	0.83	0.76	0.73
Below poverty line (%)	0.02	0.79	0.65	0.58
Median income	-0.01	0.88	0.78	0.76
Home Rule	0.05	0.55	0.38	0.33

Table 4: Standardized covariate mean differences and randomization test p-values for Complete Randomization (CR), Paired Randomization (PR), and Constrained PR for the Keele et al. (2017) matched dataset.

observed one. Table 4 shows the resulting p-values for each covariate and design. Our test indicates that all of the covariates are well-balanced according to these designs.

To determine which of these designs is most appropriate for this matched dataset, we computed the Mahalanobis distance across 10,000 draws from Complete Randomization, Paired Randomization, and Constrained Paired Randomization. Figure 3 shows the resulting distribution of Mahalanobis distances under these three designs, as well as the observed Mahalanobis distance for the matched dataset. We can see that the matched dataset is unusually well-balanced compared to what we would expect from Complete Randomization and Paired Randomization, whereas the observed balance is near the mode of what we would expect from Constrained Paired Randomization, indicating that it may be the most appropriate design for this matched dataset. The corresponding p-values are 0.97 for Complete Randomization, 0.83 for Paired Randomization, and 0.72 for Constrained Paired Randomization, again indicating the plausibility of these three designs.

5.3 Causal Analyses Under Different Experimental Designs

Now we will analyze the matched data under Complete Randomization, Paired Randomization, and Constrained Paired Randomization, all of which were found to be plausible in the previous section. As in Section 4, we will construct randomization-based confidence intervals for this matched dataset by inverting the hypothesis $H_0^\tau : Y_i(1) = Y_i(0) + \tau$ using the mean-difference estimator as the test statistic. Table 5 shows our randomization-based confidence intervals assuming Complete Randomization, Paired Randomization, and Constrained Paired Randomization, as well as the confidence interval reported in Keele et al. (2017). All of the confidence intervals suggest that the presence of at least one African American candidate in Louisiana mayoral elections significantly increases black voter turnout.

The confidence interval in Keele et al. (2017) was constructed by inverting Wilcoxon's signed rank test, which is a common approach for analyzing matched-pair data (Rosenbaum, 2002). This is why our randomization-based confidence interval assuming Paired Randomization is quite similar to the Keele et al. (2017) confidence interval. These analyses condition on the matched pairs—which is why they are substantially narrower than the complete randomization confidence interval—but they do not condition on the covariate

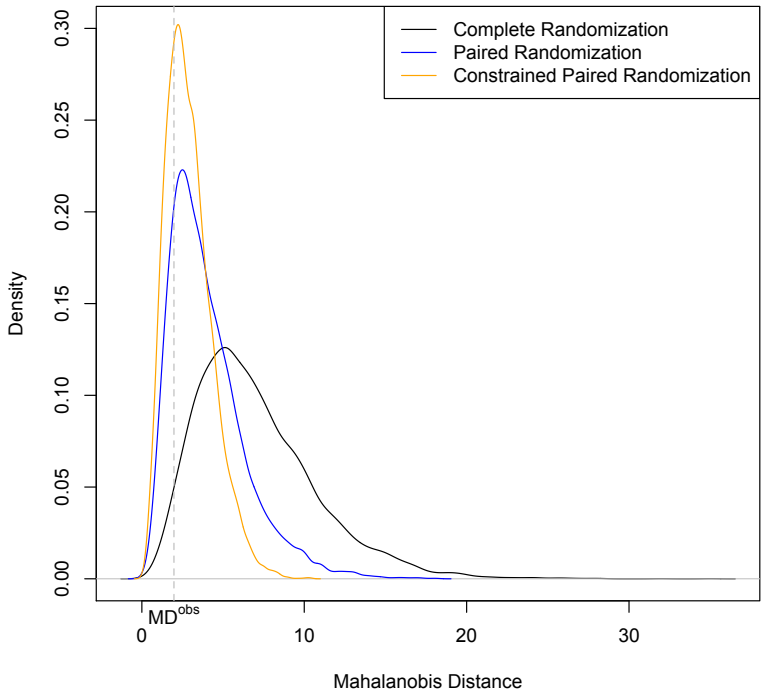


Figure 3: Distribution of the Mahalanobis distance for different designs. We generated 10,000 random draws from Complete Randomization in (1), Paired Randomization in (11), and Constrained Paired Randomization in (12); computed the Mahalanobis distance of each draw; and plotted the corresponding distributions and the observed Mahalanobis distance in the matched dataset. The corresponding p-values are 0.97 for Complete Randomization, 0.83 for Paired Randomization, and 0.72 for Constrained Paired Randomization.

Analysis	**Confidence Interval**
Keele et al.	(0.74, 6.04)
Complete Rand.	(0.09, 6.75)
Paired Rand.	(0.70 6.11)
Constrained Paired Rand.	(0.81 6.00)

Table 5: Confidence intervals for the ATE after matching. Units are in percentage points.

balance across pairs. Meanwhile, Constrained Paired Randomization does, resulting in a slightly narrower confidence interval: It is approximately 78% the width of the complete randomization confidence interval and 96% the width of the paired randomization confidence interval. This is in line with the width reduction we saw via simulation in Section 4 for the third outcome, which was only somewhat linearly related with the covariates in that simulation ($R^2 = 0.25$ on average across the 1,000 simulated datasets). In this application, when we regressed black voter turnout on all the covariates in the full dataset, $R^2 = 0.18$, and when we regressed black voter turnout on all the covariates but election year (which was almost exactly matched in each pair), $R^2 = 0.12$. The graphical diagnostic in Figure 3 provides evidence in favor of using the narrower confidence interval assuming Constrained Paired Randomization for this dataset. Even though all of the confidence intervals in Table 5 suggest the same conclusion for this application, they demonstrate how assuming different designs within an observational study can lead to different levels of precision.

However, an important contribution of Keele et al. (2017) was conducting a sensitivity analysis to assess how sensitive their results were to hidden biases; they were able to conduct such an analysis using tools that assume a paired design (e.g., Rosenbaum 2002, Chapter 4). A promising line of future work is extending these tools to more complex designs, like Constrained Paired Randomization.

6. Discussion and Conclusion

Covariate imbalance is one of the principal problems of causal inference in observational studies. To tackle this problem, matching algorithms can produce datasets with strong covariate balance. It is common to assume matched datasets approximate completely randomized or block randomized experiments if covariate balance diagnostics are met, even though these diagnostics do not formally assess whether treatment is effectively randomized. Instead, we propose a randomization test for assessing if a particular experimental design is plausible for a matched dataset. Our test is a generalization of other randomization tests for assessing covariate balance (Hansen & Bowers, 2008; Cattaneo et al., 2015; Gagnon-Bartsch & Shem-Tov, 2019), where we can test any experimental design, including designs with covariate balance constraints. In the analysis stage, we recommend a randomization-based approach, which can flexibly incorporate any assignment mechanism—a design-based decision—and any treatment effect estimator—a model-based decision.

Our test, like all balance tests, is limited in that failing to reject does not "prove" that treatment is effectively randomized or that assuming randomized assignment is guaranteed to yield valid inferential results. Nonetheless, our test serves as a helpful tool for detecting clear violations of random assignment that harm inferential results for matched data. In particular, we recommend using the Mahalanobis distance as a test statistic, because it accounts for the joint behavior of covariates while acting as a global test for balance. Meanwhile, we found that well-designed matched datasets that exhibit high levels of covariate balance tend to approximate balance-constrained designs like rerandomization. Analyzing these matched datasets as such can lead to precise causal analyses. However, assuming a precise experimental design for a matched dataset should be proceeded with caution, because it can harm inferential results if there are still imbalances in relevant covariates after matching.

To demonstrate how to use our balance diagnostics and inferential approach in practice, we revisited a causal analysis conducted in political science by Keele et al. (2017). Researchers often have field-specific knowledge guiding them towards balancing particular covariates when matching, which was the case in this application. Using our approach, we pinpointed the experimental designs the resulting matched dataset may approximate. Furthermore, we improved the precision of this causal analysis by conditioning on the high levels of balance that researchers—using subject-matter expertise—ensured by design.

Because our framework combines design-based and model-based decisions, a promising line of future research is comparing different combinations of these decisions and assessing which combinations yield the best inference for a matched dataset. In particular, our simulations suggest that there may be an equivalence between certain designs when linear regression is used. This echoes recent findings that design-based matching methods and model-based machine learning methods often yield similar results (Keele & Small, 2020). Furthermore, our approach can be applied to settings beyond matching. For example, assumptions of random assignment have been used in regression discontinuity designs (Li et al., 2015; Cattaneo et al., 2015; Mattei & Mealli, 2016; Branson & Mealli, 2018) and instrumental variable approaches (Brookhart & Schneeweiss, 2007; Baiocchi et al., 2014; Branson & Keele, 2020). Our `randChecks` `R` package—used throughout this paper—can formally test effective random assignment of any binary indicator, such as binary treatments in regression discontinuity designs and binary instrumental variables.

Acknowledgments

We would like to thank Stephen Blyth, Luis Campos, Lucas Janson, Edward Kennedy, Luke Miratrix, Reagan Mozer, Nicole Pashley, and an anonymous reviewer for insightful comments that led to substantial improvements in this work. We would also like to thank Luke Keele for providing the data used in our application. This research was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1144152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Replication Code for the Lalonde Example

Here we demonstrate how the tests and visuals presented in Section 2.4 were created using our `randChecks` `R` package. This also serves as a demonstration as to how `randChecks` can be used to assess random assignment of a binary indicator, e.g., a treatment within a matched dataset.

The `randChecks` `R` package is available on `CRAN`; after installing the package, one can load the three datasets in Section 2.4 using the following line of code:

```
data("lalondeMatches")
```

This loads three `R` objects:

- `lalonde`: The full Lalonde dataset of 614 subjects, 185 of whom are in treatment and 429 are in control.
- `lalonde.matched.ps`: The 1:1 propensity score matched dataset created using the `MatchIt` `R` package. This dataset contains 370 subjects, 185 of whom are in treatment and 185 are in control. Thus, compared to the full Lalonde dataset, this dataset contains the full treatment group but a subset of the control group.
- `lalonde.matched.card`: The 1:1 cardinality matched dataset created using the `designmatch` `R` package. This dataset was created such that all standardized covariate mean differences are below 0.1. This dataset contains 240 subjects, 120 of whom are in treatment and 120 are in control. Thus, compared to the full Lalonde dataset, this dataset contains a subset of the treatment and control groups.

Figure 1a displays a Love plot of the standardized covariate mean differences for these three datasets. To create this plot, we first defined the covariate matrix and treatment indicator for these three datasets:

```
#obtain the covariates for these datasets
X.lalonde = subset(lalonde, select = -c(treat))
X.matched.ps = subset(lalonde.matched.ps, select = -c(treat,subclass))
X.matched.card = subset(lalonde.matched.card, select = -c(treat,subclass))
#the treatment indicator is
indicator.lalonde = lalonde$treat
indicator.matched.ps = lalonde.matched.ps$treat
indicator.matched.card = lalonde.matched.card$treat
```

Then, one can use the `getStandardizedCovMeanDiffs()` function within `randChecks` to define the standardized covariate mean differences:

```
meanDiffs.lalonde = getStandardizedCovMeanDiffs(X.matched = X.lalonde,
  indicator.matched = indicator.lalonde)
meanDiffs.matched.ps = getStandardizedCovMeanDiffs(
  X.matched = X.matched.ps, indicator.matched = indicator.matched.ps,
  X.full = X.lalonde, indicator.full = indicator.lalonde)
meanDiffs.matched.card = getStandardizedCovMeanDiffs(
  X.matched = X.matched.card, indicator.matched = indicator.matched.card,
  X.full = X.lalonde, indicator.full = indicator.lalonde)
```

This allows us to create the Love plot in Figure 1a. Note that the arguments `X.full` and `indicator.full` are used such that the standardized covariate mean differences have the same denominator across all datasets.

Meanwhile, Figure 1b displays the Love plot for the 1:1 propensity score matched dataset, along with quantiles for the standardized covariate mean differences under complete randomization (i.e., permutations of the treatment indicator). This figure acts as a visualization of our randomization test using the standardized covariate mean differences as a test statistic; it was generated using the `lovePlot()` function within `randChecks`:

```
lovePlot(X.matched = X.matched.ps, indicator.matched = indicator.matched.ps,
  X.full = X.lalonde, indicator.full = indicator.lalonde,
  permQuantiles = TRUE,
  perms = 1000,
  assignment = "complete")
```

The argument `permQuantiles = TRUE` states that the quantiles across permutations should be plotted; `perms = 1000` states that 1,000 permutations should be used; `assignment = "complete"` states that the permutations should be under complete randomization.

Finally, Figure 2 displays the distribution of the Mahalanobis distance under complete randomization, paired randomization, and constrained randomization (where the standardized covariate mean differences are constrained to be less than 0.1) for the 1:1 cardinality matched dataset. This figure acts as a visualization of our randomization test using the Mahalanobis distance as a test statistic. To assess paired randomization, we needed to define the pair labeling (known as a `subclass` within the `randChecks R` package):

```
subclass.matched.card = lalonde.matched.card$subclass
```

Then, Figure 2 was generated using the `asIfRandPlot()` function within `randChecks`:

```
asIfRandPlot(X.matched = X.matched.card, indicator.matched = indicator.matched.card,
  X.full = X.lalonde, indicator.full = indicator.lalonde,
  assignment = c("complete", "blocked", "constrained diffs"),
  subclass = subclass.matched.card,
  perms = 1000,
  threshold = 0.1)
```

This function was also used to generate Figure 3 in the real data application (Section 5). Furthermore, our randomization test results in Sections 2.4, 4, and 5 were generated using the `asIfRandTest()` function within `randChecks`, which has the same arguments as the `asIfRandPlot()` function.

Appendix B. Data Generating Process for Simulations

Here we provide details about the data generating process from the simulation study in Section 4. We followed the simulation setup of Austin (2009b) and Resa & Zubizarreta (2016). We generated 1,000 datasets, where each dataset contained $N_T = 250$ treated

subjects and $N_C = 500$ control subjects. Each subject has eight covariates, generated as such:

$$\begin{aligned}(x_{i1}, x_{i2}, x_{i3}, x_{i4}) &\sim \mathcal{N}_4\left(W_i \begin{pmatrix}0.2\\0.2\\0.5\\0.5\end{pmatrix}, \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\right)\\ (x_{i5}, x_{i6}) &\sim \text{Bern}\left(0.1 + 0.068W_i\right)\\ (x_{i7}, x_{i8}) &\sim \text{Bern}\left(0.4 + 0.242W_i\right)\end{aligned} \tag{13}$$

where $W_i = 1$ if subject i is treated and 0 otherwise. This is similar to "Scenario 1" of Resa & Zubizarreta (2016); the other two scenarios consider heterogeneous variances and collinearity between the treatment and control groups, and we differ to their work for the performance of matching under those scenarios. The above covariates are generated such that the true standardized difference in means—which is $\frac{\mu_T - \mu_C}{\sqrt{\frac{\sigma_T^2 + \sigma_C^2}{2}}}$ for Normal random variables[9] and $\frac{p_T - p_C}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}}$ for Bernoulli random variables[10] (Rosenbaum & Rubin, 1985)—is 0.2 for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5$, and $\mathbf{x}_6$, and 0.5 for $\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_7$, and $\mathbf{x}_8$.

After the covariates were generated, three outcomes were generated for each subject:

$$y_{i1} = f_1(x_i) + W_i + \epsilon_i \tag{14}$$
$$y_{i2} = f_2(x_i) + W_i + \epsilon_i \tag{15}$$
$$y_{i3} = f_3(x_i) + W_i + \epsilon_i \tag{16}$$

where $\epsilon \overset{iid}{\sim} N(0, 4)$. Thus, the outcomes are generated noisily around the mean functions f_1, f_2, and f_3, with an additive treatment effect of one. The mean functions are:

$$f_1(x_i) = 3.5x_{i1} + 4.5x_{i3} + 1.5x_{i5} + 2.5x_{i7} \tag{17}$$
$$f_2(x_i) = f_1(x_i) + 2.5\text{sign}(x_{i1})\sqrt{|x_{i1}|} + 5.5x_{i3}^2 \tag{18}$$
$$f_3(x_i) = f_2(x_i) + 2.5x_{i3}x_{i7} - 4.5|x_{i1}x_{i3}^3| \tag{19}$$

Thus, the outcomes $\mathbf{y}_1, \mathbf{y}_2$, and $\mathbf{y}_3$ are ordered in terms of increasing complexity. Furthermore, only the odd-numbered covariates are included in the outcome, in order to mimic the fact that not all available covariates necessarily affect the outcome in practice (Resa & Zubizarreta, 2016).

References

Abadie, A., & Spiess, J. (2020). Robust post-matching inference. *Journal of the American Statistical Association*, (pp. 1–37).

Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade, M., Allman, R. M., Meleth, S., & Bourge, R. C. (2006). Heart failure, chronic

9. Here, μ_T and μ_C are the population-level means for the treatment and control groups, respectively, and σ_T^2 and σ_C^2 are analogously defined for the variances.
10. Here, p_T and p_C are the population-level proportions for the treatment and control groups, respectively.

diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *European Heart Journal*, *27*(12), 1431–1439.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049.

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107.

Austin, P. C. (2009b). Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and monte carlo simulations. *Biometrical Journal*, *51*(1), 171–184.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057–1069.

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340.

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. B. (2002). School choice in ny city: A bayesian analysis of an imperfect randomized experiment. In *Case Studies in Bayesian Statistics*, (pp. 3–97). Springer.

Barreto, M. A., Segura, G. M., & Woods, N. D. (2004). The mobilizing effect of majority–minority districts on latino turnout. *American Political Science Review*, *98*(1), 65–75.

Bind, M.-A. C., & Rubin, D. B. (2019). Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical Methods in Medical Research*, *28*(7), 1958–1978.

Bobo, L., & Gilliam, F. D. (1990). Race, sociopolitical participation, and black empowerment. *American Political Science Review*, *84*(2), 377–393.

Box, G. E., Hunter, W. H., & Hunter, S. (1978). *Statistics for Experimenters*, vol. 664. John Wiley and Sons New York.

Branson, Z., & Bind, M.-A. (2019). Randomization-based inference for bernoulli trial experiments and implications for observational studies. *Statistical methods in Medical Research*, *28*(5), 1378–1398.

Branson, Z., & Keele, L. (2020). Evaluating a key instrumental variable assumption using randomization tests. *American Journal of Epidemiology*, *189*(11), 1412–1420.

Branson, Z., & Mealli, F. (2018). Local randomization and beyond for regression discontinuity designs: Revisiting a causal analysis of the effects of university grants on dropout rates. *arXiv preprint arXiv:1810.02761*.

Branson, Z., & Shao, S. (2021). Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity. *Journal of Statistical Planning and Inference*, *211*, 287–314.

Brookhart, M. A., & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics*, *3*(1).

Browning, R. P., Marshall, D. R., & Tabb, D. H. (1984). *Protest is not enough: The struggle of blacks and Hispanics for equality in urban politics*. University of California Press.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, *3*(1), 1–24.

Caughey, D., Dafoe, A., & Miratix, L. (2016). Beyond the sharp null: Permutation tests actually test heterogeneous effects. In *Summer Meeting of the Society for Political Methodology, Rice University, July*, vol. 22.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199.

Dasgupta, T., Pillai, N. S., & Rubin, D. B. (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B*, *77*(4), 727–753.

Dehejia, R. H. (2003). Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data. *Journal of Business & Economic Statistics*, *21*(1), 1–11.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84*(1), 151–161.

Diamond, A., & 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*(3), 932–945.

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical Science*, *32*(3), 331–345.

Edgington, E., & Onghena, P. (2007). *Randomization Tests*. CRC Press.

Flury, B. K., & Riedwyl, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician*, *40*(3), 249–251.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, *105*(4), 994–1000.

Fogarty, C. B. (2020). Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, *115*(531), 1518–1530.

Frangakis, C. E., Rubin, D. B., & Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, *3*(2), 147–164.

Gagnon-Bartsch, J., & Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, *13*(3), 1464–1483.

Good, P. (2013). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Science & Business Media.

Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, *5*(2), 263–275.

Greifer, N. (2017). cobalt: Covariate balance tables and plots. *R package version*, *2*(0).

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618.

Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2050–2054.

Hansen, B. B. (2009). Propensity score matching to recover latent experiments: diagnostics and asymptotics. *Balance*, *322*, 4103.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*(2), 219–236.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234.

Hartman, E., & Hidalgo, F. D. (2018). An equivalence approach to balance and placebo tests. *American Journal of Political Science*, *62*(4), 1000–1013.

Heckman, J. J., Lopes, H. F., & Piatek, R. (2014). Treatment effects: A bayesian perspective. *Econometric Reviews*, *33*(1-4), 36–67.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., & Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, *4*(1), 61–80.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Ho, D. E., Imai, K., King, G., Stuart, E. A., et al. (2011). Matchit: nonparametric preprocessing for parametric causal inference. *J Stat Softw*, *42*(8), 1–28.

Hodges Jr, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, (pp. 598–611).

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, (pp. 945–960).

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345–361.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, *27*(24), 4857–4873.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*, *171*(2), 481–502.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, *21*(62), 1–54.

Keele, L., & Small, D. S. (2020). Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. *The American Statistician*, (pp. 1–25).

Keele, L. J., Shah, P. R., White, I., & Kay, K. (2017). Black candidates and black turnout: A study of viability in louisiana mayoral elections. *The Journal of Politics*, *79*(3), 780–791.

Kilcioglu, C., & Zubizarreta, J. R. (2016). Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. *The Annals of Applied Statistics*, *10*(4), 1997–2020.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, (pp. 604–620).

Lee, W.-S. (2013). Propensity score matching and variations on the balancing test. *Empirical Economics*, (pp. 1–34).

Leighley, J. E. (2001). *Strength in numbers?: The political mobilization of racial and ethnic minorities*. Princeton University Press.

Li, F., Mattei, A., & Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, *9*(4), 1906–1931.

Li, F., Thomas, L. E., & Li, F. (2018a). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, *188*(1), 250–257.

Li, X., & Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Li, X., Ding, P., & Rubin, D. B. (2018b). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, *115*(37), 9157–9162.

Li, X., Ding, P., & Rubin, D. B. (2020). Rerandomization in 2^k factorial experiments. *The Annals of Statistics*, *48*(1), 43–63.

Linden, A., & Yarnold, P. R. (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, *22*(6), 848–854.

Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, *65*(1), 21–30.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*(456), 1245–1253.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49–55.

Mattei, A., & Mealli, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies*, (2), 156–173.

Ming, K., & Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, *10*(3), 455–463.

Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B*, *75*(2), 369–396.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263–1282.

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

Pashley, N. E., & Miratrix, L. W. (2017). Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv:1710.10342*.

Pimentel, S. D., Kelz, R. R., Silber, J. H., & 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*(510), 515–527.

Ramsahai, R. R., Grieve, R., & Sekhon, J. S. (2011). Extending iterative matching methods: an approach to improving covariate balance that allows prioritisation. *Health Services and Outcomes Research Methodology*, *11*(3-4), 95–114.

Resa, M., & Zubizarreta, J. R. (2016). Evaluation of subset matching methods and forms of covariate balance. *Statistics in Medicine*, *35*(27), 4961–4979.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation*.

Rosenbaum, P. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. (2010). *Design of Observational Studies*. New York: Springer.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B*, (pp. 597–610).

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B*, *67*(4), 515–530.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, *21*(1), 57–71.

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

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

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, (pp. 159–183).

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, (pp. 34–58).

Rubin, D. B. (1980). Comment on 'Randomization analysis of experimental data: The Fisher randomization test' by Debabrata Basu. *Journal of the American Statistical Association*, *75*(371), 591–593.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36.

Rubin, D. B. (2008a). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484), 1350–1353.

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, (pp. 808–840).

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585.

Samii, C., & Aronow, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, *82*(2), 365–370.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software, Forthcoming*.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*, 487–508.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1-2), 305–353.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, *101*(476), 1398–1407.

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of "A critical appraisal of propensity score matching in the medical literature between 1996 and 2003" by peter austin, statistics in medicine. *Statistics in Medicine*, *27*(12), 2062–2065.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1.

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8).

Tchetgen, E. J. T., & VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, *21*(1), 55–75.

Vegetabile, B. G., Gillen, D. L., & Stern, H. S. (2019). Optimally balanced gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Wang, Y., & Zubizarreta, J. R. (2019). Large sample properties of matching for balance. *arXiv preprint arXiv:1905.11386*.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, *47*(2), 965–993.

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*(500), 1360–1371.

Zubizarreta, J. R., & Keele, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association*, *112*(518), 547–560.

Zubizarreta, J. R., & Kilcioglu, C. (2016). Designmatch: Construction of optimally matched samples for randomized experiments and observational studies that are balanced by design. *R package version 0.1*, *1*, 187.

Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, (pp. 204–231).