Covariate-adaptive randomization inference in matched designs

Samuel D. Pimentel and Yaxuan Huang [*]

March 21, 2024

Abstract

It is common to conduct causal inference in matched observational studies by proceeding as though treatment assignments within matched sets are assigned uniformly at random and using this distribution as the basis for inference. This approach ignores observed discrepancies in matched sets that may be consequential for the distribution of treatment, which are succinctly captured by within-set differences in the propensity score. We address this problem via covariate-adaptive randomization inference, which modifies the permutation probabilities to vary with estimated propensity score discrepancies and avoids requirements to exclude matched pairs or model an outcome variable. We show that the test achieves type I error control arbitrarily close to the nominal level when large samples are available for propensity score estimation. We characterize the large-sample behavior of the new randomization test for a difference-in-means estimator of a constant additive effect. We also show that existing methods of sensitivity analysis generalize effectively to covariate-adaptive randomization inference. Finally, we evaluate the empirical value of combining matching and covariate-adaptive randomization procedures using simulations and analyses of genetic damage among welders and right-heart catheterization in surgical patients. **Keywords:** causal inference, matching, permutation test, propensity score, sensitivity analysis.

1 Introduction

1.1 Re-evaluating a common model for inference

Randomized trials provide an effective means for measuring the effect of a treatment of interest relative to a control condition for at least two reasons. First, random allocation of treatment to units ensures that large differences in pre-treatment characteristics between the group of units selected for treatment and the group of units selected for control arise in large samples only with very small probability. Second, the known distribution of indicators of treatment across units in the study provides a basis for inference. By considering each unit's potential outcome values under treatment and under control as fixed latent values and repeatedly permuting treatment labels across study units (or using tests that rely on the large-sample behavior of such permutations) one may obtain inferences without making strong assumptions about the sampling procedure used to select the study units or the model for the outcome variable. This conceptual approach, dating back at least to Fisher (1935), is often known as randomization inference.

In contrast, in observational studies of a binary effect concerns arise about whether units receiving treatment and units receiving control are otherwise comparable. If there is confounding, or systematic differences in variables (either observed or unobserved) that are predictive of the outcome of interest, the effect estimate from a simple group comparison will generally differ systematically from the effect that would

[*] Samuel D. Pimentel is an Assistant Professor and Yaxuan Huang is a doctoral student in the Statistics Department at University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720. Correspondence should be addressed to spi@berkeley.edu. Samuel D. Pimentel is supported by the National Science Foundation under Grant No. 2142146, and both authors are supported by the Hellman Fellows Program. Part of this research was performed while Samuel D. Pimentel was visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation (Grant No. DMS-1929348). The authors would like to thank Avi Feller, Kevin Guo, Ben Hansen, and Ruoqi Yu for helpful comments.

have been measured in a randomized trial. To adjust for observed confounding variables, researchers may estimate an outcome model in the absence of treatment and compare units with similar expected outcomes under control, estimate a treatment model or propensity score and compare units with similar propensities to receive treatment, or some combination of the two. Matched observational studies adjust for observed confounding by grouping each treated unit with one or more similar control units and excluding controls not sufficiently similar to any treated unit (Stuart 2010). When matching is conducted without replacement of controls so that matched sets are disjoint, and with exact agreement on a propensity score so that units grouped together shared identical propensities for treatment, then each matched set is like a miniature randomized trial; conditional on one unit within the group receiving treatment, each is uniformly likely to have been the one selected. As such, methods of randomization inference are frequently applied to matched observational studies as though in a stratified randomized trial (Silber et al. 2020; Jain et al. 2022; Shin 2022; Tesema et al. 2023). In the language of Zhang and Zhao (2023), the resulting procedures may be denoted "quasi-randomization tests."

Randomization inference in randomized trials depends on exact knowledge of the true randomization probabilities, and use of these methods in matched studies is motivated by an ideal setting in which the true propensity scores are known and matched exactly. In reality, however, propensity scores must be estimated, and except in cases where the measured variables are few and discrete they are never matched exactly. Optimal matching procedures such as those described in Rosenbaum (1989), Zubizarreta (2012), Austin and Stuart (2015), and Pimentel et al. (2020) use estimated propensity score differences as important inputs, so that when a match is created the researcher has information easily available about which matched sets are relatively closer or further from achieving the ideal uniform distribution for treatment assignment. Yet this information is not used when it comes time to do inference. By using uniform randomization inference, researchers implicitly assume a much simpler model and hope that these differences are all small enough not to create substantial lack of fit. While sensitivity analyses for unobserved confounding that are often conducted post hoc can in principle subsume discrepancies in observed variables too, these analyses are not typically presented in these terms, nor are observed differences typically used to calibrate the parameters for these sensitivity analyses.

We propose a new method for inference in matched observational studies, covariate-adaptive randomization inference, that explicitly uses estimated propensity score discrepancies to update permutation probabilities. This approach retains most of the advantages of uniform randomization inference – clear conceptual connections to a hypothetical randomized trial, ease of implementation, compatibility with interpretable methods of sensitivity analysis — while addressing potential for lack of fit even in settings where propensity score differences need not disappear in large samples. Furthermore, the lack of fit adjustment generally does not require users to alter the match itself in ways that reduce overall sample size. Although covariate-adaptive randomization inference can in theory fully resolve the confounding problem that matching itself seeks to address, we also demonstrate that matching followed by covariate-adaptive randomization offers precision and robustness benefits not enjoyed by covariate-adaptive randomization in an unmatched study.

1.2 Related work

Covariate-adaptive randomization inference builds on a quickly-growing literature that explores using estimated propensity scores to structure permutation tests. Rosenbaum (1984) recommended permuting treatment assignment conditional on the sufficient statistic for a propensity score fit, a closely related idea which works very well for settings with only one or two discrete covariates with a limited number of categories.

In a graduate dissertation Baiocchi (2011) briefly proposed permuting treatment assignments within matched pairs in a manner similar to that described above, although with a slightly different distribution based on a ratio of propensities on the probability scale rather than the odds scale. More recently, Branson and Bind (2019) describe randomization tests with non-uniform treatment assignment probabilities for unmatched Bernoulli trials; although they focus on randomized trials they also suggest the use of the method for observational studies. The conditional permutation test of Berrett et al. (2020) proposes using permutations of observed variables based on an estimated conditional distribution for the purposes of testing conditional independence. Shaikh and Toulis (2021) also use permutations of observations based on estimated propensity scores to conduct causal inference in an observational study and obtain even stronger guarantees about large-sample performance by leveraging specific focus on a setting in which only one unit receives treatment. Resampling observations with non-uniform probabilities is also an important component of modern conformal inference for settings where data are not exchangeable (Tibshirani et al. 2019), including observational studies of treatment effects (Lei and Candès 2021). Our contribution to this literature is to articulate the particular advantages of combining non-uniform permutation with matched designs. The structure of matched sets helps resolve many of the practical challenges that arise in conducting non-uniform permutation tests, such as the question of how to sample from the permutation distribution discussed at length by Berrett et al. (2020). In addition, covariate-adaptive randomization inference in concert with matching has important precision benefits relative to covariate-adaptive randomization in an unmatched study, as we will document in Section 6. In contrast to previous authors, we also integrate covariate-adaptive randomization inference into the sensitivity analysis framework of Rosenbaum (2002b), allowing us to conduct inference in the presence of unobserved confounding variables.

At a high level, covariate-adaptive randomization inference may also be compared to inverse probability weighting approaches to causal inference. Both methods rely on estimated propensity scores to adjust for observed confounding variables, but the two approaches use the estimated scores differently. In inverse weighting, estimators for a treatment effect are constructed by dividing each observed outcome by the estimated probability of its observed treatment (or some function thereof) before comparing average outcomes across groups. This corrects potential outcomes in expectation and produces unbiased estimates of average treatment effects (Mukerjee et al. 2018), but can also lead to high-variance estimators since inverse weights can be large when propensity scores are near zero or one (Robins and Wang 2000; Kang and Schafer 2007). In contrast, covariate-adaptive randomization inference uses estimated propensity scores to correct the null distribution of an arbitrary test statistic rather than to construct a specific test statistic Of course, inverse weighting estimators may also be used within the covariate-adaptive randomization inference framework, an idea we explore in Section A.1 of the online appendix.

Another difference between the inverse weighting framework and covariate-adaptive randomization inference is the type of null hypothesis typically tested. Weighting approaches usually test weak null hypotheses that only place restrictions on average potential outcomes, remaining mostly agnostic to how treatment effects vary across individuals, while covariate-adaptive randomization inference as introduced here focuses on sharp null hypotheses that define a specific treatment effect for each individual (Ding and Dasgupta 2016). While weak-null approaches account naturally for treatment effect heterogeneity, they also rely on asymptotic distributions of test statistics while a covariate-adaptive test of a sharp null does not (only requiring asymptotic convergence of the propensity score estimate to the true propensity score, a potentially weaker assumption when the propensity score is estimated in a separate dataset as discussed in Section 4 below). Weak null approaches generally require more restrictive conditions on test statistics than permutation tests of the sharp null, which are valid for arbitrary functions of treatment and outcome. Finally, an extensive sensitivity analysis framework has been developed for tests of the sharp null (Rosenbaum 2002b), which we adapt below in Section 5. Note also that tests designed for sharp null hypotheses can also be adapted to test weak nulls instead (Caughey et al. 2021; Fogarty 2022), a point we discuss further in Section 8.

1.3 Outline

In what follows we develop and evaluate covariate-adaptive randomization inference. In Section 2 we introduce a formal framework and describe the shortcomings of uniform randomization inference in detail. In Section 3 we introduce covariate-adaptive randomization inference, giving procedures for hypothesis testing and for estimation, and confidence interval construction under a constant additive effect model. In Section 4 we bound the error introduced into the procedure by estimation of the propensity score. In Section 5, we generalize sensitivity analysis procedures grounded in uniform randomization inference to covariate-adaptive randomization inference to allow valid inference in the presence of unobserved confounding variables. In Section 6, we demonstrate finite sample performance of covariate-adaptive hypothesis tests and confidence intervals and compare to alternative strategies, including both matching approaches that permute subjects uniformly and unmatched approaches that use a form of covariate-adapative randomization inference. In Section 7 we demonstrate implications for practice by reanalysis of two observational datasets: one measuring genetic damage experienced by welders and one assessing the impact of right-heart catheterization on patient mortality. Finally, Section 8 highlights important questions and connections raised by this work and outlines opportunities for further research.

2 Formal framework and problem setup

2.1 Uniform randomization inference in matched designs

Consider a population of individuals each represented by a vector $(Y(1), Y(0), Z, X, U)$. Z is a binary indicator for membership in the treatment group, X is a vector of observed covariates, and U is an unobserved covariate. The values $Y(z)$ are potential outcomes under different treatment conditions as defined under the Neyman-Rubin causal model and the stable unit treatment value assumption (Rubin 1980; Holland 1986), which specifies that an individual's outcome depends only on its own treatment status, rather than on the treatment status of other individuals, and that the only versions of treatment are 0 and 1. Only the information (Y, Z, X) is observed by the analyst, where Y represents the observed outcome $Y(Z)$. Let $\lambda(\mathbf{x}) = P(Z|X = x)$ be the conditional probability of treatment given observed covariates, or the propensity score, and let $\pi(x, u) = P(Z|X = x, U = u)$ be the true probability of treatment (which also depends on the unobserved U).

We assume individuals are first sampled independently from the population, and then formed into a matched design $\mathcal{M}$, consisting of K matched sets, on the basis of their treatment variables Z and covariates X alone. Each matched set, numbered $k = 1$ through K, contains exactly one treated individual and one or more control individuals. Individuals in set k are numbered $k1$ through kn_k, where n_k is the number of units in set k, in arbitrary order, so that we may refer to the treatment indicator of the ith unit in the kth matched set as Z_{ki}. We also define $n = \sum_{k=1}^{K} n_k$. Let vectors $\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{Y}, \mathbf{Z}, \mathbf{U} \in \mathbb{R}^n$ contain the observed univariate data ordered so that units in the same matched set are contiguous. Let matrix $\mathbf{X}$ contain all n vectors X_{ki} in its rows with an identical ordering; in addition, we abuse notation slightly to let $\lambda(\mathbf{X})$ represent the n-vector of true propensity scores. A matched design is said to be exact on a particular variable or quantity $\mathbf{v} \in \mathbb{R}^n$ if for any matched set k, $v_{ki} = v_{kj}$ for all $i, j \in \{1, \ldots, n_k\}$. Let $\mathcal{Z}_{\mathcal{M}}$ be the set of all treatment vectors $\mathbf{Z}'$ such that $\sum_{i=1}^{n_k} Z'_{ki} = 1$ for all matched sets k.

We now consider the the distribution of treatment assignments conditional on the match selected and the potential outcomes, i.e $P(\mathbf{Z} \mid \mathcal{F})$ where $\mathcal{F} = \{\mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\}$. Rosenbaum (2002b, §3) argues that this distribution is discrete uniform on $\mathcal{Z}_{\mathcal{M}}$ under two key assumptions: first, the absence of unobserved confounding, under which $\lambda(x) = \pi(x, u)$ for all u, and exact matching on covariates $\mathbf{X}$ (or more generally exact matching on true propensity scores $\lambda(X)$). The null distribution for a test statistic T can then be derived under the following sharp null hypothesis of no treatment effects for any individual in the sample:

$$\mathbf{Y}(1) = \mathbf{Y}(0). \tag{1}$$

In particular, the sharp null hypothesis of no effect guarantees that the actual observed outcome Y_{ki} for individual i in matched set k would still have been observed had Z_{ki} taken on a different value. Thus under the sharp null the test statistic $T(Z, Y)$ varies conditional on $\mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)$ only through the vector $\mathbf{Z}$, which is uniformly distributed over all permutations $\mathbf{Z}_{perm}$ of the observed element of $\mathbf{Z}$ within matched sets. The exact p-value for the test of the sharp null, where we reject for larger values of $T(Z, Y)$, can be computed as the proportion of values of $\mathbf{Z}_{perm}$ for which $T(Z_{perm}, Y)$ exceeds $T(Z, Y)$. In practice this quantity can be computed via repeated Monte Carlo draws from the permutation distribution or via a normal approximation to the permutation distribution. For example, Rosenbaum (2002b)[§2] gives asymptotic distributions for a variety of rank statistics.

Estimates of causal effects and confidence intervals can also be developed from the null distribution associated with a sharp null hypothesis. Under a treatment effect model such as the constant additive model, the randomization test can be inverted to produce Hodges-Lehmann point estimates and corresponding confidence intervals. Alternatively, when the test statistic itself is an estimator for an effect of interest (as in the case of the difference-in-means estimator), confidence intervals can be obtained under a normal approximation in large samples by using a variance estimate obtained from the null distribution and quantiles of the normal probability density function.

2.2 Propensity score discrepancies imperil the uniform treatment distribution

While the assumption of exact matching on a true propensity score yields convenient mathematical symmetry matching is never exact on all measured covariates in datasets of any substantial scale, nor is it practical to match exactly even on a univariate true propensity score, both because propensity scores must be estimated in practice and because they are often modeled as smooth functions of continuous variables that do not agree exactly for any two subjects. As such it is important to consider the potential for resulting lack of fit between the nominal uniform distribution of treatment used for inference and the true distribution for a given matched design. Hansen (2009) addressed this question for the difference-in-means estimator of an additive treatment effect and found potential for slow-shrinking finite sample bias when the true propensity score is not matched exactly. Other authors have shown that bias in treatment effect estimation (Sävje 2021) and failure of Type I error control for the uniform randomization test (Guo and Rothenhäusler 2023) persist even in infinite samples in settings where all treated units are matched in pairs without replacement except in unusual situations such as populations where the probability of propensity scores exceeding 0.5 is zero or the true outcome model is known. Of course, even bigger problems may arise if unobserved confounding is also present, but methods of sensitivity analysis have been constructed with specific attention to this issue, as will be explored in greater detail in Section 5.1.

It has been argued that certain kinds of balance tests may be understood as tests for lack of fit between the actual distribution of treatment and the uniform model of inference used (Hansen 2009), so that if balance tests do not reveal problems then this problem can be ignored. This approach falls short of resolving the problem optimally for two reasons. First and most importantly, the theory underlying these balance tests relies on an asymptotic regime in which propensity score differences within matched sets approach zero as sample size increases, which in turn comes from a pattern of ever-larger concentrations of control units in a region arbitrarily close to any given treated unit. In many settings, this assumption is not reasonable. For example, Pimentel et al. (2015) consider the common setting under which matches are conducted within natural blocks or groups of units, such as matching patients within hospitals or students within schools. When, as in these examples, the size of an individual block may reasonably be viewed as bounded and the most natural way to think about increasing sample size is by adding more blocks, there is no reason to expect propensity score differences within matched sets to shrink to zero, since the concentration of matchable controls near a given treated unit is limited by the upper bound on the size of a block. More generally, Sävje (2021) demonstrated that when propensity scores larger than 0.5 are present in the population with probability exceeding 0 and matching is conducted without replacement, then some matched sets will necessarily have propensity score discrepancies bounded away from zero. A second problem with the balance testing solution is that it does not fully articulate how to resolve problems with a matched design when the balance test fails. Common solutions such as using a tighter propensity score caliper lead to tradeoffs by reducing other aspects of match quality such as the proportion of treated units retained.

Another proposed strategy is the use of regression adjustment to remove slow-shrinking bias not addressed by close matching on a propensity score or on covariates (Abadie and Imbens 2011; Guo and Rothenhäusler 2023). Under assumptions on the outcome model, this method removes the bias, and under sufficiently strong assumptions on the outcome model and the convergence of matched discrepancies to zero, a fast rate is achieved. However, the assumptions may not always be plausible, particularly the condition that the propensity score discrepancies shrink to zero as discussed by Sävje (2021). In addition, estimating an outcome model may be inconvenient if the outcome is multivariate, is related to observed covariates in a complex or poorly-understood manner, or is not yet measured at the time the match is conducted.

2.3 Lack of fit due to dependence between the true treatment vector and the match itself

All of the above discussion focuses on discrepancies between the uniform distribution of treatment given $\mathcal{F} = \{\mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\}$ and the actual distribution of treatment given these quantities, thinking of the match $\mathcal{M}$ as fixed over all possible $\mathbf{Z}$-values. However, if a model of independently-sampled subjects is assumed on the original data prior to matching then one might view the matched design, which is constructed with reference both to $\mathbf{X}$ and $\mathbf{Z}$, as a random function of treatment. In the special case of exact matching this issue need not arise, as the specific match chosen may remain conditionally independent of $\mathbf{Z}$ given event $\mathcal{Z}_{\mathcal{M}}$. However, when matching is not exact this guarantee no longer holds, and for some $\mathbf{Z} \in \mathcal{Z}_{\mathcal{M}}$ it may be the case that match $\mathcal{M}$ never would have been selected. For example, a treated unit with a high propensity score may be difficult to match and be paired with a control having an appreciably lower propensity score; however, if the treated unit had been a control instead, it would have been possible to form a better match with a treated unit having a similarly high propensity score value. Values of $\mathbf{Z}$ that would have produced a different match do not belong in the support of the distribution of treatment conditional on $\mathcal{M}$, but they appear with positive support in the other two distributions mentioned. The phenomenon of reduced support will be denoted Z-dependence (with reference to the dependence of the match chosen on the original treatment status vector).

This issue is also discussed by Pashley et al. (2021), who note, “proper conditional analysis would need to take into account the matching algorithm." However, no such analysis has yet been proposed in the matching literature, and Pashley et al. suggest that matching may perform well empirically even in the presence of Z-dependence. Z-dependence is not a central focus in what follows, and unless otherwise noted we will take the perspective that the matched sets $\mathcal{M}$ are fixed. However,in Section 6.4, we explore a "proper conditional analysis" that does account for the matching algorithm in a very small simulated dataset using brute-force methods, and find potential for empirically meaningful type I error violations. Z-dependence will also be helpful in explaining the empirical performance of uniform and covariate-adaptive randomization inference in certain larger simulation settings considered in Sections 6.2-6.3. We advocate and outline ideas for further progress in understanding Z-dependence in Section 8.

3 Adapting randomization inference to covariate discrepancies

3.1 True and estimated conditional distributions of treatment status

To adapt randomization inference to discrepancies in observed covariates, we begin by considering the case where no unobserved confounding is present and represent the true conditional distribution in terms of propensity scores. Since treatment indicator Z_{ki} is Bernoulli($\lambda(X_{ki})$), we have:

$$\begin{aligned}
&P\left\{Z_{ki} = 1 \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\right\} \\
&= \frac{P\left\{Z_{ki} = 1, Z_{k2} = 0, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k)\right\}}{\sum_{j=1}^{n_k} P\left\{Z_{k1} = 0, \ldots Z_{k(j-1)} = 0, Z_{kj} = 1, Z_{k(j+1)} = 0, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k)\right\}} \\
&= \frac{\lambda(X_{ki}) \prod_{j \neq i}[1 - \lambda(X_{kj})]}{\sum_{j=1}^{n_k} \lambda(X_{kj}) \prod_{\ell \neq j}[1 - \lambda(X_{kj\ell}))]} = \frac{\text{odds}\{\lambda(X_{ki})\}}{\sum_{j=1}^{n_k} \text{odds}\{\lambda(X_{kj})\}} = p_{ki}.
\end{aligned}$$

Because individuals are sampled independently, the joint conditional probability for a treatment vector $\mathbf{Z}$ is given by multiplying the appropriate p_{ki} terms together:

$$P\left\{\mathbf{Z}_{\mathcal{M}} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\right\} = \begin{cases} \prod_{k=1}^{K} \prod_{i=1}^{n_k} p_{ki}^{z_{ki}} & \mathbf{z} \in \mathcal{Z}_{\mathcal{M}} \\ 0 & \mathbf{z} \notin \mathcal{Z}_{\mathcal{M}} \end{cases} \tag{2}$$

If the true propensity score $\lambda(\cdot)$ were known, the p_{ki}s could be calculated exactly. To conduct inference under the sharp null, the Monte Carlo strategy described in Section 2.1 could be used, except that instead of permuting treatment assignments within matched sets uniformly at random one would draw the identity of the treated unit in each set from an independent multinomial random variable with 1 trial and probabilities $p_{k1}, \ldots, p_{kn_k}$. Exact finite-sample confidence intervals could also be obtained by inverting the test, and the large majority of the benefits of the uniform randomization inference procedure could be retained despite the reality of inexact matching on the propensity score. Note that when propensity scores are matched exactly, this procedure reduces to the uniform test.

Unfortunately, the propensity score is typically unknown, so the true conditional distribution of treatment is also unknown. However, it is straightforward to construct a plugin estimator for this distribution by substituting an estimate of the propensity score $\widehat{\lambda}(\cdot)$ for $\lambda(\cdot)$ in the formula for p_{ki}. We denote the process of conducting hypothesis tests and creating confidence intervals using this plugin distribution as covariate-adaptive randomization inference. In concrete terms, an investigator may conduct covariate-adaptive randomization inference via the Monte Carlo approach of Section 2.1 by calculating the estimated propensity odds for each subject in a matched set, by dividing each estimated odds by the sum of the odds in the matched set, by determining treatment via a single draw from a multinomial distribution with probabilities for each subject given by these normalized odds, and by conducting this process independently within all matched sets, repeating as necessary to generate a close approximation to the null distribution.

3.2 The difference-in-means estimator: large-sample distribution

While taking repeated Monte Carlo draws from the covariate-adaptive permutation distribution is straightforward, it is also instructive to construct a normal approximation using the mean and variance of the test statistic. Here we demonstrate this approach for the standard difference-in-means estimator, defined below (for brief discussion of an alternative estimator based on inverse propensity weighting, see Section A.1 of the online appendix).

$$T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) = \frac{1}{K}\sum_{k=1}^{K}\left\{\sum_{i=1}^{n_k} Z_{ki}Y_{ki} - \frac{1}{n_k-1}\sum_{i=1}^{n_k}(1-Z_{ki})Y_{ki}\right\} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_kZ_{ki}-1}{n_k-1}.$$

Under the sharp null hypothesis, the $\mathbf{Y}$ is fixed across all values for $\mathbf{Z}$ so the expectation and the variance are calculated only over the random variables $\mathbf{Z}$.

$$E(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}(1), \mathbf{Y}_{\mathcal{M}}(0)) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_kp_{ki}-1}{n_k-1} \tag{3}$$

$$Var(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}(1), \mathbf{Y}_{\mathcal{M}}(0)) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\frac{n_k}{n_k-1}\right)^2 Y_{ki}p_{ki}\left\{Y_{ki}(1-p_{ki}) - \sum_{j\neq i}^{n_k} Y_{kj}p_{kj}\right\}$$

Note that these formulas get considerably simpler in the case of a matched pair design, in which $n_k = 2$ for all k. Here the inner sum $\sum_{i=1}^{n_k} Y_{ki}\frac{n_kZ_{ki}-1}{n_k-1}$ may be rewritten as $(Z_{k1} - Z_{k2})(Y_{k1} - Y_{k2})$, so that the mean of the test statistic is the sample average (across matched pairs k) of V_kD_k where $V_k = p_{k1} - p_{k2}$ and $D_k = Y_{k1} - Y_{k2}$, and the variance is the sample average of $K^{-1}(1 - V_k^2)D_k^2$.

While this mean and variance of the test statistic depend on the unknown true propensity score through the p_{ki} terms, they can be estimated by substituting the $\widehat{p}_{ki}$ obtained by substituting the estimated propensity score $\widehat{\lambda}$ for the true propensity score λ. Large-sample inference may be conducted by computing a normalized deviate of the difference-in-mean statistic using estimates of the mean and variance above, and comparing to the standard normal distribution. This relies on a finite-sample central limit theorem based on a sequence of infinitely-expanding finite samples (Li and Ding 2017) for which the mean and variance of the test statistic converge to stable limits. The most natural asymptotic regime places some upper bound on the size of a matched set n_k across all these samples while K grows towards infinity, since returns to matching large numbers of controls to a single treated unit are quickly diminishing (Hansen 2004). The simplest available central limit theorem is thus in the inner sums $\sum_{i=1}^{n_k} Y_{ki}\frac{n_kZ_{ki}-1}{n_k-1}$, which are independent random variables with non-identical distributions; when n_k is uniformly bounded and a Lindeberg condition holds on the potential outcomes $Y(0)$, they follow from results in Liu and Yang (2020). For more discussion of asymptotic regimes in matching with restricted or fixed matching ratios, and of central limit theorems for settings with many small strata, see Abadie and Imbens (2006) and Liu et al. (2022).

The large-sample approximation just described does not explicitly account for random variation due to the estimation of the propensity score; in particular, the mean term, which is a function of $\widehat{p}_{ki}$ random variables, is instead treated as fixed. To probe the possible role of such variation, we used the M-estimation framework (Stefanski and Boos 2002) to represent both the propensity score estimates and the test statistic of interest as part of a single multivariate estimation problem and to construct a new variance estimate for the difference between the test statistic and its estimated mean that attempts to incorporate variation in the estimated propensity scores. However, the new variance estimates were frequently near-identical or slightly smaller than those ignoring the propensity score estimation (presumably due to positive correlation between the estimated mean and the test statistic across propensity score fits); in our simulations the associated tests performed almost identically. Given the simpler form and intuitive permutation analogue for the test ignoring propensity score estimation, we focus on this test going forward. More details are provided on the M-estimation approach in Section A.4 of the online appendix.

3.3 Estimates and confidence intervals for an additive effect

So far we have discussed covariate-adaptive randomization inference primarily through the lens of testing the sharp null hypothesis of zero effect. It is also possible to use covariate-adaptive randomization inference to construct confidence intervals in settings where the primary goal is estimating a specific causal effect. The exact details of this process depend on the estimand; for purposes of exposition, we focus on estimating a constant additive treatment effect. More formally, we assume that there exists τ such that

$$Y_{ki}(1) = Y_{ki}(0) + \tau \qquad \text{for all } i. \tag{4}$$

and we seek to estimate τ and obtain a confidence interval for our estimate.

First, note that for $\tau \neq 0$, the sharp null hypothesis of no effect no longer holds and observed outcomes Y_{ki} are no longer invariant to treatment status, but the transformed outcomes $Y_{ki} - Z_{ki}\tau$ will be invariant, and a randomization test can be constructed by using $Y_{ki} - Z_{ki}\tau$ as input to the test statistic rather than Y_{ki}. This adjustment allows us to test $H_0 : \tau = \tau_0$ for any value $\tau_0 \in \mathbb{R}$. The set of τ_0-values for which the corresponding test does not reject H_0 at level α is a $1 - \alpha$ confidence set for τ (Neyman 1937; Lehmann 1959, §3.5).

To obtain an estimate of τ in model (4), we may simply select the value of τ_0 with the largest two-sided p-value (or the median of such values if many share the same p-value). This estimator was previously suggested for use in treatment-control studies by Branson and Bind (2019), who noted its connection to the Hodges-Lehmann estimator (Hodges Jr and Lehmann 1963); for a helpful review of Hodges-Lehmann estimation in the context of randomization inference in observational studies, see Rosenbaum (2002b, §2.7.2). Coudin and Dufour (2020) also show that in a slightly different context, an estimator based on maximizing the p-value of a permutation tests retains many of the attractive theoretical properties of the traditional Hodges-Lehmann estimator, including median unbiasedness and asymptotic normality, under mild regularity assumptions. Numerical simulations show that this estimator generally performs well and consistently improves substantially on the naïve difference-in-means estimator (see Table 7 in Section A.5 of the online supplement).

Because covariate-adaptive randomization inference induces a discrete distribution for the test statistic, some care is require in inverting the tests to obtain confidence intervals, especially in small samples. In particular, when the distribution's discreteness makes it impossible to construct an interval with coverage exactly $1 - \alpha$, it is important to ensure that intervals are made conservative rather than anticonservative. Luo et al. (2021) provide a very helpful practical discussion; while they focus on a uniform design without stratification, the key ideas translate almost without modification to covariate-adaptive randomization inference at least when the difference-in-means statistic of Section 3.2 is used. Inverting the test requires repeated tests for a variety of candidate τ-values and can be computationally demanding. However, several shortcuts are possible. In matched pair designs, Monte Carlo draws of the difference-in-means statistic from the null distribution under $H_0 : \tau = \tau_0$ are simple algebraic modifications of Monte Carlo draws from the null distribution under $H_0 : \tau = 0$. In particular, letting Z_{ki}^{orig} represent the original value of treatment in the observed data for subject i in pair k (as opposed to the value Z_{ki} in the draw from the null distribution), we have the following:

$$\begin{aligned} T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}} - \mathbf{Z}_{\mathcal{M}}^{orig}\tau_0) &= \frac{1}{K}\sum_{k=1}^{K}(Z_{k1} - Z_{k2})[Y_{k1} - Y_{k2} - \tau_0(Z_{k1}^{orig} - Z_{k2}^{orig})] \\ &= T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) - \tau_0 \cdot \frac{1}{K}\sum_{k=1}^{K}(Z_{k1} - Z_{k2})(Z_{k1}^{orig} - Z_{k2}^{orig}) \\ &= T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) - \tau_0 \cdot \frac{1}{K}\left[\sum_{k=1}^{K}\mathbf{1}\left\{\mathbf{Z}_k = \mathbf{Z}_k^{orig}\right\} - \sum_{k=1}^{K}\mathbf{1}\left\{\mathbf{Z}_k \neq \mathbf{Z}_k^{orig}\right\}\right] \end{aligned}$$

In summary, as long as we keep track of the number of pairs in each null draw that have been switched relative to the original treatment assignment, we need only take Monte Carlo draws once from the null distribution for $H_0 : \tau = 0$ and can simply transform them as necessary to test any $H_0 : \tau = \tau_0$. Another option not limited to paired designs is to use the large-sample normal approximation to the null distribution of the difference-in-means estimator instead of Monte Carlo draws from the permutation distribution. To invert this large-sample test, one need only evaluate a normal tail probability for each τ_0 considered.

4 Assessing the impact of propensity score estimation error

A primary concern in determining the empirical value of covariate-adaptive randomization inference is understanding the impact of errors in estimating the propensity score. To trust the method, we need some guarantee that small deviations between $\widehat{\lambda}(x)$ and true $\lambda(x)$ values result in only small deviations in nominal and actual test size and confidence interval coverage. The following results provide partial reassurance in this regard. To lay the groundwork, let $\widehat{\lambda}_N(\cdot)$ be an estimated propensity score fit on an external sample of size N from the infinite population, and let the quantities $\mathbf{Z}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)$ refer to a separate sample organized into K matched pairs containing $\sum_{k=1}^{K} n_k = n$ total units. Let F_λ represent the true conditional distribution of $\mathbf{Z}$ as a function of the true propensity score $\lambda(\cdot)$ and let $F_{\widehat{\lambda}_N}$ represent the distribution of $\mathbf{Z}$ used to conduct covariate-adaptive inference in practice, based on $\widehat{\lambda}_N(\cdot)$. Furthermore, let $p_{\widehat{\lambda}_N,n}$ be the p-value produced by a nominal level-α covariate-adaptive randomization test of the sharp null hypothesis using $F_{\widehat{\lambda}_N}$. The following result, adapted from Berrett et al. (2020) (who consider the slightly different case in which permutations are across the entire dataset and not within matched sets), relates these quantities using the total variation distance of F_λ and $F_{\widehat{\lambda}_N}$.

Theorem 1. *If no unobserved confounding is present and the sharp null hypothesis of no effect is true, then*

$$P(p_{\widehat{\lambda}_N,n} \leq \alpha) \ \leq \ \alpha + d_{TV}\left(F_\lambda, F_{\widehat{\lambda}}\right)$$

where $d_{TV}(P, Q)$ gives the total variation distance between two probability distributions P and Q.

In words, this theorem says that the true type I error of a covariate-adaptive randomization test performed with nominal level α is no larger than α plus the total-variation discrepancy of the distributions implied by the true and estimated propensity score. Intuitively, when the estimated propensity score is close to the true propensity score, this means that the nominal type I error rate is close to the true type I error rate. We formalize this intuition for the case in which the true propensity score obeys a logistic regression model.

Theorem 2. *Suppose that $P(Z = 1 \mid X) = \lambda(X) = \frac{1}{1+\exp(-\beta^T X)}$ and that $\widehat{\lambda}_N$ is obtained by estimating this model using maximum likelihood. Suppose furthermore that N increases with the primary sample size n such that $\lim_{n \longrightarrow \infty} n/N = 0$, and suppose the covariates X have compact support. Then under the conditions of Theorem 1,*

$$\limsup_{n,N \longrightarrow \infty} P(p_{\widehat{\lambda}_N,n} \leq \alpha) \ \leq \ \alpha.$$

The proof, which uses Pinsker's inequality to bound the total variation distance by a sum of Kullback-Leibler divergences and a Taylor expansion to show that this sum converges to zero, is deferred to Section A.2 of the online appendix.

A natural question is whether similar results can be obtained when the propensity model is not necessarily logistic. Berrett et al. (2020) sketch a proof for a result similar to Theorem 2 when the propensity score is estimated nonparametrically using a kernel method; this requires only mild smoothness conditions rather than a correctly-specified model, although the bound on the rate of convergence of the size of the Type I error violation towards zero is much weaker. More generally, one may turn to Theorem 1 directly to explore misspecification: this result provides a bound on Type I error violation whenever the degree of misspecification can be quantified by the total variation distance between estimated and true conditional distributions of treatment. Such metrics for probing robustness appear elsewhere in the literature; for example, Guo et al. (2022) assess robustness of inferences to covariate noise by considering a regime in which the conditional distribution of the true covariates given treatment has bounded total variation distance from the conditional distribution of the noise-contaminated covariates.

Another limitation of Theorem 2, which extends also to the nonparametric kernel approach just mentioned, is the asymptotic regime. The issues extend beyond the usual question of whether a particular sample size is large enough to ensure that error is small; here the requirement that the pilot sample used to estimate the propensity score grow at a larger rate than the analysis sample raises similar concerns even for very large samples, since the key question is whether the pilot sample's size sufficiently exceeds that of the analysis sample. Fitting the propensity score in a large independent sample consisting of most of the observed data while reserving a relatively smaller portion for the analysis is uncommon in applied matching studies. However, we note that this approach is natural in settings where treatment and covariate values are readily available but outcomes are expensive or difficult to measure, as when researchers must collect outcomes by administering a test to study subjects (Reinisch et al. 1995) or by abstracting medical charts (Silber et al. 2001). In Section 6, we evaluate the performance of the method by simulation in multiple settings, including some that plausibly adhere to the assumed asymptotic regime and others that likely do not. In general we find only minor difference in performance for tests based on estimated propensity scores fit out-of-sample in samples much larger than the analysis samples versus tests based on propensity scores estimated in-sample on the analysis data, suggesting that the test may be fairly robust to violations of the asymptotic regime given in Theorem 2 in many finite sample settings.

5 Covariate-adaptive randomization inference under unobserved confounding

5.1 Review of sensitivity analysis framework under exact matching

Results in Sections 2 -4 have all depended on the absence of unobserved confounding, but in practice it is not plausible that all confounders are observed. To address this issue with the matched randomization inference framework, Rosenbaum (2002b, §4) presents a method of sensitivity analysis to relax the no unobserved confounding assumption. Specifically, the true probabilities of treatment are restricted as follows:

$$1/\Gamma \leq \frac{\pi(x,u)(1-\pi(x,u'))}{\pi(x,u')(1-\pi(x,u))} \leq \Gamma \quad \text{for all } x, u, u'. \tag{5}$$

This is equivalent to the following model for treatment assignment, with arbitrary $\kappa(\cdot)$:

$$\log\left(\frac{\pi(X,U)}{1-\pi(X,U)}\right) = \kappa(X) + \gamma U \quad \text{where } \gamma = \log(\Gamma) \text{ and } 0 \leq U \leq 1. \tag{6}$$

To complete the sensitivity analysis, some method is needed to compute worst-case p-values over all possible values of of $\mathbf{U}$ allowed by the model. The method depends on the structure of matched sets formed and the test statistic used; we focus on the approach of Rosenbaum (2018) which allows for arbitrary numbers of controls matched to each treated unit and applies to any sum statistic, i.e. any statistic $T(\mathbf{Z}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y})$ for chosen functions f_{ki} of $\mathbf{Y}$, which we denote as the pseudo-responses. As shown in Rosenbaum (2018, §5), the difference-in-means statistic, the regression-adjusted test statistics of Rosenbaum (2002a), and other M-statistics are all members of the family of sum statistics.

Without loss of generality, suppose the n_k units in each matched set k are arranged in increasing order of their pseudo-responses f_{ki} so $f_{k1} \leq f_{k2} \leq \ldots, f_{kn_k}$ for all k, and suppose that we are interested in a one-sided test where larger values of the test statistic will lead to rejection. Let U^+ be the set of all stratified N-tuples $\mathbf{u}$ such that $u_{ki} \in \{0,1\}$ and $u_{k1} \leq u_{k2} \leq \ldots \leq u_{kn_k}$ for all k, and let U^- be the set of all stratified N-tuples $\mathbf{u}$ such that $u_{ki} \in \{0,1\}$ and $u_{k1} \geq u_{k2} \geq \ldots \geq u_{kn_k}$ for all k. Let $\alpha_{unif}(\mathbf{Y}, \mathbf{u})$ represent the p-value for the uniform randomization test performed with test statistic $T(\mathbf{Z}, \mathbf{Y})$ and the uniform randomization distribution when model (6) holds with unobserved confounder $\mathbf{U} = \mathbf{u}$. When matching on the propensity score is exact, Rosenbaum and Krieger (1990) showed the following:

$$\min_{\mathbf{u} \in U^-} \alpha_{unif}(\mathbf{Y}, \mathbf{u}) \leq \alpha_{unif}(\mathbf{Y}, \mathbf{U}) \leq \max_{\mathbf{u} \in U^+} \alpha_{unif}(\mathbf{Y}, \mathbf{u}).$$

Although $\mathbf{U}$ is still unknown, this statement allows us to bound its impact on the result of the hypothesis test by searching over a highly structured finite set of candidate $\mathbf{u}$-values. We now show that this approach still works to identify the worst-case p-value when matches are not exact on the propensity score and permutation probabilities are covariate-adaptive.

5.2 Sensitivity analysis model under covariate-adaptive randomization inference.

As in the previous section, let $T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})$ be a sum statistic and focus on the case of a one-sided test where large values lead to rejection. We now define $\alpha_{adapt}(\mathbf{Y}, \mathbf{u})$ as the p-value obtained by conducting a covariate-adaptive randomization test (using the true propensity score) when model (6) holds with unobserved confounder $\mathbf{U} = \mathbf{u}$.

Theorem 3. *For any* $\mathbf{u} \in [0,1]^n$, $\min_{\mathbf{u}' \in U^-} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}') \leq \alpha_{adapt}(\mathbf{Y}, \mathbf{u}) \leq \max_{\mathbf{u}'' \in U^+} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}'')$.

Intuitively, this result says that we can find the maximum (minimum) p-value by placing the maximum (minimum) possible treatment probability on the subjects with the largest pseudo-responses, and the minimum (maximum) treatment probability on those with the smallest pseudo-responses. The full proof, which is partially adapted from results for the exact-matching case in Rosenbaum (2002b) and from results for weighting estimators in Zhao et al. (2019), is deferred to Section A.3 of the online appendix.

While Theorem 3 provides a foundation for sensitivity analysis, two additional issues must be addressed. First, while in the exact-matching case the function $\kappa(X)$ has no bearing on sensitivity analysis (since the $\kappa(X)$ terms cancel within matched pairs), under covariate-adaptive randomization inference some bound or estimate of $\kappa(X)$ is needed to compute sensitivity bounds. In general, the known propensity score $\lambda(X)$ provides information about $\kappa(X)$ as follows:

$$\lambda(X) = E(\pi(X,U) \mid X) = \int_0^1 \pi(X,u)dP(u) = \int_0^1 \frac{1}{1+\exp[-\kappa(X) - \gamma u]}dP(u).$$

When $\gamma = 0$ this gives us a one-to-one mapping between $\kappa(X)$ and $\lambda(X)$ but otherwise such a mapping requires knowledge about the population distribution of U. Fortunately, since $\pi(x,u)$ is increasing in u for all x, for any fixed x we have $\pi(x,0) \leq \lambda(x) \leq \pi(x,1)$. This in turn implies:

$$\kappa(X) \in \left[\log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - \gamma, \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right)\right]$$

Note that this bound is tight, since we can make $\kappa(X)$ arbitrarily close to either bound by letting $P(U=0)$ or $P(U=1)$ be arbitrarily close to 1. Therefore we can rewrite model (6) in terms of $\lambda(X)$ as:

$$\begin{aligned}\log\left(\frac{\pi(X,U)}{1-\pi(X,U)}\right) &= \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - \gamma V + \gamma U, \qquad U, V \in [0,1]. \\ &= \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - 1 + 2\gamma U', \qquad U' \in [0,1]. \qquad (7)\end{aligned}$$

The second issue is computational. The bound in Theorem 3 gives us a finite set of possible distributions over which to search to identify the worst-case p-value for a covariate-adaptive randomization test. However, under certain configurations of strata size and number, this set may grow large and complicated so that it is difficult to compute the exact maximum efficiently. Gastwirth et al. (2000) simplified the problem by showing that asymptotically the overall maximum and minimum p-values for any given Γ are achieved by solving simpler optimization problems separately for each stratum and aggregating the results. Specifically, if $T_k = \sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y}_\mathcal{M})$ $\mu_{k\ell}$ is the additive contribution of stratum k to the test statistic and $\mathbf{u}_k$ is the subvector of $\mathbf{u}$ for stratum k, the expectation of T_k must be maximized (minimized) over all binary $\mathbf{u}_k$, and when multiple values $\mathbf{u}_k$ are maximizers (minimizers) the variance of T_k must also be maximized over these. By an argument essentially identical to the one we use to prove Theorem 3, one may show that T_k's expectation is maximized only for $\mathbf{u}_k$ such that $u_{k1} = \ldots = u_{k\ell} = 0$ and $u_{k(\ell+1)} = \ldots, u_{kn_k} = 1$ for some $\ell = 0, 1, \ldots, n_k$, so that it suffices to consider the n_k expectations $\mu_{k\ell}$ and n_k variances $\nu_{k\ell}$ associated with these vectors (for the lower bound similar quantities are used but with a different choice of binary u_{ki} values). Aggregating worst-case expectations and variances across many strata and letting the number of strata go to infinity results in worst-case bounds on the overall p-value. While this result is asymptotic, Rosenbaum (2018) later derived a one-step adjustment that renders the bounds conservative in finite samples. Under covariate-adaptive randomization inference and representation (7) for the treatment model, the key quantities $\mu_{k\ell}$ and $\nu_{k\ell}$ take on the following values:

$$\mu_{k\ell} = \frac{\sum_{i=1}^{\ell} p_{ki} f_{ki} + \Gamma^2 \sum_{\ell+1}^{n_k} p_{ki} f_{ki}}{\sum_{i=1}^{\ell} p_{ki} + \Gamma^2 \sum_{\ell+1}^{n_k} p_{ki}} \qquad \nu_{k\ell} = \frac{\sum_{i=1}^{\ell} p_{ki} f_{ki}^2 + \Gamma^2 \sum_{\ell+1}^{n_k} p_{ki} f_{ki}^2}{\sum_{i=1}^{\ell} p_{ki} + \Gamma^2 \sum_{\ell+1}^{n_k} p_{ki}} - \mu_{k\ell}^2$$

The method for combining these quantities to provide an overall conservative bound on the p-value is identical to that described in Rosenbaum (2018).

6 Finite-sample performance via simulation

6.1 Motivation

The results of Section 4 provide some assurance that in extremely large samples, covariate-adaptive randomization inference will closely approximate the true conditional distribution of treatment and outperform uniform randomization inference. However, many matched studies work with relatively small sample sizes, and researchers do not always have strong guarantees that fitted models are well-specified. In what follows we assess the empirical performance of covariate-adaptive randomization inference in small to moderate samples via a simulation, considering cases where models are correctly and incorrectly specified. First, in Section 6.2, we consider the performance of covariate-adaptive randomization inference in matched designs compared to more traditional inference procedures for matched studies, with a focus on Type I error control.

A second important question is whether modifying a permutation procedure to account for observed propensity scores obviates the need for matching at all. Indeed, the test proposed by Branson and Bind (2019), which considers a similar data setting and relies on the same idea as our test but does not incorporate matching, is also finite-sample valid when conducted with true propensity scores. In Section 6.3, we conduct additional simulations to demonstrate that the combination of matching and covariate-adaptive randomization inference offers important benefits in terms of both precision and robustness to misspecification of the propensity model relative to non-uniform permutation in an unmatched study.

6.2 Matching with and without covariate-adaptive randomization inference

We create a matrix of covariates X by drawing p vectors of independent standard normal random variables, each vector of length n. p is 2, 5, or 10, and n is either 100 or 1000. The true propensity score is then given by one of the following two functions, one that specifies the logit of treatment as a linear function of the columns of X and another specifying it as a nonlinear function of those columns:

$$\text{logit}\left[P(Z=1 \mid X)\right] = \log\left(\frac{0.3}{0.7}\right) + \Delta \cdot X_1 \tag{8}$$

$$\text{logit}\left[P(Z=1 \mid X)\right] = \log\left(\frac{0.3}{0.7}\right) + \frac{\Delta}{\sqrt{265}}\left(X_1 + 4X_1^3\right) \tag{9}$$

The functional forms of these models are adapted from those used by Resa and Zubizarreta (2016). The parameter Δ controls the strength of the propensity score signal; we examine two signals, “weak signal” with $\Delta = 0.2$ and “strong signal” with $\Delta = 0.6$. The intercept is chosen to ensure a treated:control ratio in the general neighborhood of 3:7 (since all X_j variables have mean zero over repeated samples), providing a relatively large control pool for pair matching. The scaling factor $1/\sqrt{265}$ is chosen to render the overall signal-to-noise ratio comparable across models for a given value of Δ. For a more detailed look at the distribution of the propensity scores in the treated and control groups under this setup, see Figure 4 in Section A.5 of the online appendix.

Matching is conducted on the joint (X, Z) datasets created by this process. The matching is done using a robust Mahalanobis distance on the columns of X, either with or without a propensity score caliper. When the caliper is used, a propensity score must first be estimated. This is done by fitting a logistic model with linear additive terms using maximum likelihood. Then matches are restricted to occur only between individuals separated by no more than 0.2 sample standard deviations of the fitted propensity scores (computed for the original dataset); if treated units must be excluded from the match to meet this condition the minimal possible number of such exclusions is made.

Finally, outcomes are drawn using a linear model with the same right-hand side as the model used to generate the true propensity score, albeit with no intercept, with $\Delta = 1$, and with additive independent mean-zero normal errors with variance 4. We conduct randomization inference by reshuffling treatment indicators within pairs uniformly at random, based on propensity scores estimated as described above, or based on true propensity scores. The observed test statistic is compared to the null distribution obtained by 5000 Monte Carlo draws to see whether one-sided tests for a difference greater than zero reject at the 0.05 level. Results are evaluated for both raw outcomes and outcomes adjusted by ordinary least-squares linear regression using the approach of Rosenbaum (2002a).

In addition, for each combination of simulation parameters we also test the results of inference when the observed test statistic is not generated by the original $\mathbf{Z}$-vector but by a within-match permutation of treatment indicators using distribution (2) with the true propensity scores. This is designed to approximate a model in which matched pairs have independent treatment assignments, and eliminates the potential for Z-dependence as discussed in Section 2.3.

For datasets of size $n = 100, p = 2$ and $n = 100, p = 5$ we ran each unique combination of the remaining simulation parameters (propensity model, signal strength, caliper indicator, inference method, regression adjustment indicator, and indicator for reshuffled Z-vector in observed test statistic) 8000 times. Type I error rates are calculated as the sample average of rejection indicators over the 8000 draws for each parameter combination. We also ran each combination of parameters 8000 times for datasets of size $n = 1000, p = 10$. We opted not to examine the $n = 100, p = 10$ case because we wanted to focus on cases in which completed matches met standard balance criteria and sample runs suggested that most matches were not successful in balancing all ten covariates, and we studied only $p = 10$ for the much more computationally expensive $n = 1000$ case in order to produce simulation results in a reasonable timeframe.

Figure 1 shows the primary results on Type I error rate from the simulations. The most apparent pattern is the high type I error rates for uniform inference on uncalipered matches with the difference-in-means statistic, much higher than the rates for any method that accounts in any additional way for in-pair covariate discrepancies. This is true across every simulation setting shown but especially so at $n = 1000$.

More nuanced are the distinctions among settings using covariate-adaptive inference, calipers, regression-adjusted test statistics, and combinations thereof. In the first three columns of each table, corresponding to settings in which Z-dependence is present, settings with calipers or regression adjusted test statistics tend to perform best, either with or without covariate-adaptive inference; in contrast settings relying entirely on covariate-adaptive inference tend to do slightly worse or even substantially worse when $n = 1000$. However, when treatment assignments within matched pairs are made independent, eliminating Z-dependence, covariate-adaptive inference alone is generally just as effective as regression adjustment and does better than calipers.

Across the four horizontal groupings in each table, we see different combinations of correctly/incorrectly specified treatment and outcome models. Covariate-adaptive inference with estimated propensity scores tends to fare more poorly when the treatment model is misspecified, even when Z-dependence is not present. Regression adjustment is somewhat robust to the degree of nonlinearity introduced in this simulation's outcome model. However, when both outcome and treatment models are nonlinear type I error control often fails for $p > 2$. At $n = 1000$ the case with both models misspecified leads to gross violation of type I control under any form of adjustment except oracle covariate-adaptive inference (which uses true, nonlinear treatment probabilities that are unavailable in practice).

All the results shown above focus on the strong propensity signal case ($\Delta = 0.6$), in which matching tends to be substantially more difficult than in the weak signal case. Density plots and type I error rates for the weak-signal case are given in Figures 5 and 6 in Section A.5 of the online appendix. While the overall ordering among approaches remains similar, the size of type I error violations is greatly reduced such that the simple uniform approach often controls Type I error when $p = 2$.

We also conducted two robustness checks to confirm that our results are not sensitive to secondary aspects of our simulation setting. Firstly, we reproduced the Type I error results excluding all simulation runs in which a matched sample failed to achieve absolute standardized differences less than 0.2 on all measured covariates (in case the initial simulations were contaminated by poor matches in some iterations). Secondly, we repeated our simulations with propensity scores fit on independent pilot samples of size 10,000, simulated from the same data-generating process as the analysis sample (a setting more like the one assumed in Theorem 2). These robustness checks all substantiated the patterns observed in the primary simulations (see Figures 7-8 in Section A.5 of the online appendix).

Figure 2 describes confidence interval width for simulation settings that achieved approximate Type I control. Whenever a Bonferroni-corrected test against the null hypothesis of a true Type I error rate no more than 0.05 failed to reject, we used the central limit theorem approximation of Section 3.2 to create large-sample confidence intervals for an additive effect. For cases with regression-adjusted test statistics, the procedures of Section 3.2 were used with residuals from the regression model in place of raw outcomes Y_{ki}. The two main takeaways are the value of regression adjustment for improving precision, a pattern that appears in other parts of the causal inference literature (Rosenbaum 2002a; Fogarty 2018; Antonelli et al. 2018), and a cost in precision associated with imposing calipers. This cost comes from reduced sample size, since treated units may be excluded by the caliper leading to fewer matched sets formed. Notably, introducing covariate-adaptive inference never hurt precision among cases with controlled Type I error.

In summary, the simulations emphasize the importance of taking measures of some kind beyond optimal matching itself to control covariate discrepancies within matched pairs and guarantee Type I error control for post-match inference. Moreover, the importance of such measures appears to increase with the size and complexity of the dataset. With respect to Type I error control, the type of correction or adjustment employed seems to matter less than the fact of employing it. Secondly, while covariate-adaptive inference in isolation is slightly less successful than calipers or regression adjustment in many settings, this appears to be largely a product of Z-dependence, since drawing treatment assignments independently across matched pairs tends to erase this gap in most cases. This finding suggests the importance of the specific stochastic model used to justify matched randomization inference, and the value of further work on understanding Z-dependence and finding effective ways to remove it. Finally, the simulations show that among adjustment strategies that do not rely on fitting some form of outcome model, covariate-adaptive inference holds a small edge over caliper approaches because it does not require a reduction in sample size to fix problems with in-pair covariate discrepancies.

6.3 Covariate-adaptive randomization inference with and without matching

We next conduct simulations to articulate the possible benefits of conducting matching followed by covariate-adaptive randomization relative to applying randomization inference directly to unmatched data as in Branson and Bind (2019). First we consider gains to precision in inference. The results of Section 3.2 show that the null variance of the difference-in-means statistic under covariate-adaptive randomization inference in a pair-matched study is a function of terms D_k^2 where D_k is the difference between the potential outcomes under control in pair k. While the variance of the difference-in-means statistic is much harder to compute under covariate-adaptive randomization inference under the Bernoulli assignment mechanism of Branson and Bind (2019), it tends to depend on the overall dispersion of all outcomes in the study (rather than on the dispersion within matched pairs). As a result, when heterogeneity in the study outcomes is increased but the quality of the pairings remains good, we expect to see bigger gaps in performance between matched and unmatched settings. Accordingly, for this comparison we adopt the "strong signal" version of the simulation setting in Section 6.2 but further strengthen the outcome model by multiplying the true signal by a factor of ten before adding noise. We then compare covariate-adaptive randomization inference (with and without z-dependence) to unmatched covariate-adaptive randomization inference as detailed by Branson and Bind (2019), assuming a Bernoulli design with independent treatment probabilities equal to the propensity scores. For each design, we examine both the difference-in-means and regression-adjusted test statistic.

Table 1 gives Type I errors and confidence interval lengths for the matched and unmatched versions across several parameter combinations (focusing on the case with $n = 100$, $p = 2$, strong propensity signal, and propensity scores estimated in-sample). When well-specified regression models are fit, the matched and unmatched studies show similar performance in the absence of Z-dependence, with slightly more precise confidence intervals coming from the unmatched study; this makes sense because the unmatched study has more data available with which to estimate the outcome model. However, when regression adjustment is absent or based on an incorrect model, the matched design has substantially increased precision relative to the unmatched designs, with confidence intervals less than half as large in some cases. We note also that the Type I errors for the unmatched cases without correctly-specified regression adjustment tend to lie orders of magnitude below 0.05, suggesting that the inference achieved in these cases is punishingly conservative. Evidently, matching offers potential for important precision gains relative to unmatched studies (at the cost of introducing z-dependence) even when both employ covariate-adaptive inference.

Finally, we probe the possibility that matching confers robustness to incorrect model assumptions using a new simulation setting with a badly misspecified propensity score. The covariate matrix X is generated identically to previous simulations, by drawing p vectors of independent standard normal random variables, but the true propensity score here is given by

$$\operatorname{logit}\left[P(Z = 1 \mid X)\right] = \log\left(\frac{0.2}{0.7}\right) + 1\{X_1 X_2 \geq 0\}, \tag{10}$$

Because of the indicator function on the sign of an interaction term X_1X_2, two very distinct values of the propensity score are present, associated with subjects in different quadrants of (X_1, X_2) space, but each covariate is marginally uncorrelated with treatment. The outcomes are drawn from a linear model with a zero intercept, additive noise, and the same function of covariates on the right-hand side. Table 2 presents the Type 1 errors for both the matched and unmatched version under the misspecified model (with $n = 100$, $p = 2$, and propensity scores estimated in-sample). The unmatched study exhibits uncontrolled Type I errors due to poor propensity score estimation. However, in the matched setting, Type I errors remain close to 0.05 because matched pairs with small covariate distances tend also to group subjects in the same quadrant of the (X_1, X_2) space, leading to closely-matched propensity scores. Clearly for at least some settings matching confers robustness benefits on covariate-adaptive randomization inference not enjoyed by unmatched studies.

6.4 Removing Z-dependence: a small example

The simulations of Sections 6.2-6.3 showed that Z-dependence poses a problem for settings where the original observations are sampled independently from a population prior to matching. One may argue that matching itself is best understood as a preprocessing or data cleaning step and that the model of independent assignments in matched pairs is as plausible as the model of independently sampling of observations (Ho et al. 2007), and under such a setting the results presented above without Z-dependence are most relevant for guiding practice. However, a method that can address Z-dependence under independent sampling of observations seems desirable.

With unlimited computing power, it would be trivial to construct such a method. For each draw from the null distribution, one could take the newly-generated vector of treatment assignments $\mathbf{Z}^*$ and repeat the original matching algorithm using it. If the same set of matched pairs is obtained, then $\mathbf{Z}^*$ is a valid draw from the conditional distribution of treatment given $\mathcal{Z}_\mathcal{M}$; on the other hand, if a different set of matched pairs is obtained then $\mathbf{Z}^*$ is outside the support of the true conditional distribution.

We apply this procedure in an example generated from the same data-generating process described for the unmatched vs. matched comparisons in Section 6.3 but with $n = 30$, $p = 2$, and true propensity scores. In the particular random instance we study, there are 10 treated subjects, and $2^{10} = 1024$ unique permutations of treatment assignment in the resulting ten matched pairs. Rerunning the matching algorithm with each of these permutations in turn, we find that only 300 lead to the same match. Those rejected looked systematically different than those included, with an average difference-in-means values of -0.40 vs. 1.22 respectively. The original covariate-adaptive test produces a p-value of 0.044 while the reduced-support test gives a p-value of 0.133 which is no longer significant at the 0.05 level. Repeating this analysis a total of 2000 times, we find the estimated Type I error of the original covariate-adaptive procedure to be 0.311 compared to 0.0421 in the reduced-support case. Unfortunately, this strategy does not scale easily to more realistically-sized datasets due to the explosion of the number of possible permutations and the need to compute a match for each. For more discussion, see Section 8.

7 Performance in case studies

7.1 Welders and genetic damage

We now apply the tools developed for covariate-adaptive randomization inference to a real dataset due originally to Costa et al. (1993). This dataset compares the rate of DNA-protein cross-links, a risk factor for gene expression problems, among welders and controls in other occupations. In addition to DNA-protein linkage, age, race, and smoking behavior are measured for all subjects (each of whom is male). Rosenbaum (2010a) analyzed this data by matching each of the 21 welders to one of the 26 controls. Here we replicate this matching approach and consider the impact of covariate-adaptive randomization inference in place of the uniform randomization inference typically used for pair-matched designs.

Following Rosenbaum, we estimate the propensity score using logistic regression against the three covariates discussed above and match each welder to a control using a robust Mahalanobis distance with a propensity score caliper equal to half the standard deviation of the fitted propensity score values across the dataset. Since it is not possible to match all 21 welders within the caliper, a soft caliper is used in which violations of the caliper are penalized linearly.

As noted by Rosenbaum, pair matching cannot remove all confounding in this data, particularly because only five of the 26 controls are removed by the matching process. Table 3 shows the pair differences on each of the three matching variables and the estimated propensity score, with average discrepancies at the bottom. Note that although the average discrepancies are small on all variables (the average difference in age is only 1.5 years), in the large majority of pairs the treated unit has a slightly higher propensity score, several with differences larger than 0.1. In contrast to standard uniform randomization inference, covariate-adaptive randomization inference will pay attention to these differences.

Using the difference in mean DNA-protein linkage between welders and matched controls as a test statistic, we contrast uniform randomization inference structured by the matched pairs with covariate-adaptive inference based on the estimated propensity score. The density plot in Figure 3 illustrates the difference between the two null distributions. As the consistent positive sign on the propensity score discrepancies would suggest, the covariate-adaptive randomization distribution is shifted to the right of the uniform randomization distribution, showing how accounting for residual propensity score differences leads us to expect larger values even under the null.

The data was collected based on a hypothesis that welders experience elevated levels of genetic damage compared to controls, so we conduct a one-sided test of the Fisher sharp null hypothesis of no difference in DNA-protein linkage due to welder status. Using the null distributions shown above, this corresponds to calculating the proportion of null draws that exceed the observed value of 0.64. For the uniform distribution this is 0.015, and for the covariate-adaptive distribution it is 0.029. The covariate-adaptive distribution adjusts for the residual propensity score differences in the pairs, which are biased towards treatment for the welders who actually experienced treatment; as such, it recognizes that part of the apparent treatment effect in the uniform test is likely a result of bias and concludes that the weight of evidence in favor of a treatment is weaker. While both tests are significant at the 0.05 level in this case, if researchers had been interested in effects of both signs and had conducted a two-sided test the covariate-adaptive test does not reject in this case. As such, failure to account properly for propensity score discrepancies in the two-sided case leads to an anticonservative result in which the null hypothesis is rejected when most properly it should not be. When a sensitivity analysis is run, the uniform analysis is similarly more optimistic about evidence for a treatment effect, reporting similar qualitative results for Γ up to 1.09, while the covariate-adaptive analysis allows a maximum value of only 1.05.

7.2 Right-heart catheterization and death after surgery

We next conduct covariate-adaptive randomization inference in the right-heart catheterization data of Connors et al. (1996). In this study the effectiveness of the right-heart catheterization (RHC) procedure for improving outcomes of critically ill patients was assessed by matching patients receiving the procedure to those not receiving it and comparing mortality rates. We follow the original authors by fitting a propensity score on 31 variables measured at intake and forming matches only between patients with propensity scores that differ by no more than 0.03. Unlike the authors, who use a greedy matching procedure, we use an optimal matching procedure that minimizes a robust Mahalanobis distance formed from the 31 variables in the propensity score model within propensity score calipers, and also matches exactly on primary disease category. Not all of the 2,184 RHC patients can be matched to distinct controls within the caliper, and in this case we exclude the minimal number of RHC patients necessary for the caliper to be respected. This leaves us with 1,538 matched pairs, which is still a substantial improvement on the match conducted by the authors, which had only 1,008 pairs. Table 4 summarizes the effectiveness of the match in removing bias on the observed pre-treatment variables. The numbers shown are standardized treatment-control differences in means for each variable, both before (first column) and after (second column) matching. Only the variables with the 25 largest pre-matching absolute standardized differences are shown (but post-matching standardized differences remain below 0.04 for all variables not shown). Clearly, although large differences between groups are present in the raw data for numerous variables, notably APACHE risk score (aps1), mean blood pressure (meanbp1), and P/F oxygen ratio (pafi1), matching transforms these into small differences. Of special note is the row showing balance on the estimated propensity score, which shows a reduction from a standardized difference exceeding 1 to a value of only 0.04.

To assess the role of the caliper in influencing the study's results, we construct an alternative matched design using a generalized version of optimal subset matching (Rosenbaum 2012), as implemented in the R package `rcbsubset`. Instead of using a propensity score caliper, this match enforces exact balance on ventiles of the propensity score, ensuring close balance on the propensity score without imposing hard restrictions on the propensity score discrepancy within a matched pair. Treated units are excluded from the match if their inclusion induces the average Mahalanobis distance across matched pairs to cross a specific threshold given by a tuning parameter; we chose a value of the tuning parameter that induces a similar overall sample size as in the calipered match (1,507 matched pairs). Balance on important pre-treatment variables is similar to that in the caliper match, as shown in the third column of Table 4. However, there are important differences between the two matches at the level of the pairs, as shown in Table 5; the caliper match achieves much greater similarity of propensity scores within pairs, at the cost of achieving slightly reduced similarity on a range of other variables. In practice the caliper match is likely to be more attractive, since it achieves a major improvement in propensity score control at the cost of relatively minor changes in other variables, but both matches are Pareto optimal in the sense of Pimentel and Kelz (2020).

The outcome of interest in this study is patient death within thirty days. For each match we conduct four outcome analyses: uniform and covariate-adaptive randomization inference for the difference in means, and uniform and covariate-adaptive randomization inference for a regression-adjusted difference-in-means statistic (Rosenbaum 2002a). Although the outcome is binary, the mortality rate is sufficiently high that an ordinary least-squares fit is reasonable. Table 6 summarizes the results of the analysis, providing point estimates, p-values, and estimated confidence intervals from all four approaches, with the point estimates for the covariate-adaptive procedures shifted by the mean of the covariate-adaptive null distribution given in formula (3). In addition, results from a sensitivity analysis are reported, giving the largest values of Γ consistent with a rejection of the null hypothesis (the "threshold" Γ).

Results are generally consistent with an increased risk of mortality associated with right-heart catheterization, with all confidence intervals contained in the range 3%-11%. As in the welders data, the covariate-adaptive procedure reports a treatment effect slightly smaller than the uniform procedure; however, in this case the magnitude of the difference is small compared to the overall size of the effect so the methods all agree qualitatively. Notice also that the regression-adjusted test statistics have narrower confidence intervals than the raw risk differences, in line with principles described in Rosenbaum (2002a). The regression-adjusted analyses lead to smaller estimates of the treatment effect which explains their greater sensitivity to unmeasured bias. The caliper matching has more stable threshold Γ values across analyses, consistent with the tighter control of the propensity score it achieves. For the subset match, the covariate-adaptive procedure tends to increase the threshold Γ despite having smaller treatment effects; this is because it also lowers the variance of the estimator and leads to tighter confidence intervals.

8 Discussion

Uniform randomization inference for matched studies relies, often at least in part implicitly, on a model in which unobserved confounders are absent, in which propensity scores are matched exactly, and (depending on the sampling model) in which the matched sets selected are conditionally independent of the original treatment vector. Substantial failures of these assumptions lead to substantial problems with Type I error control and require some form of correction. Covariate-adaptive randomization provides such correction by altering permutation probabilities in the randomization test based on estimated propensity scores. Relative to the naïve analysis for the difference-in-means statistic, covariate-adaptive randomization tends to restore approximate control of Type I error in many settings and constitutes an attractive option alongside approaches based on regression-adjusted test statistics and matching calipers. Furthermore, the combination of matching and covariate-adaptive randomization inference offers value not provided by incorporating estimated propensity scores into permutation procedures for unmatched studies. Specifically, matched designs can enjoy greatly improved precision relative to unmatched studies with similar non-uniform permutation procedures, and are also more robust to misspecification of the propensity score, although they are subject to an additional source of Type I error violations (z-dependence) to which unmatched studies are not. We note also that the sensitivity analysis guarantees we develop in Section 5 for matched designs do not appear to extend easily to unmatched studies, due to the greater complexity of the conditional distribution of the treatment variable.

The applied examples show that covariate-adaptive randomization inference need not change the qualitative results of an observational study, especially when the study designer has given careful attention to propensity score discrepancies. In these settings the covariate-adaptive procedure can still be a productive robustness check to help build confidence that lingering propensity score discrepancies are not corrupting the study's key findings.

Covariate-adaptive inference raises several interesting future directions for theory and methods development. First, the theoretical guarantee given in this work is limited both by a strong restriction on the relative sample sizes of the analysis sample and of a hypothetical pilot sample used to fit the propensity score. Clarifying whether this assumption can be relaxed and whether same-sample estimation of some kind, such as cross-fitting, could be applied instead would be a valuable contribution. For the difference-in-means estimator, one likely challenge is that the bias of the covariate-adaptive randomization distribution based on the estimated propensity score may not shrink to zero at a strictly faster rate than the variance when propensity-fitting and analysis samples are of a similar order, so that preserving valid inference may require widening confidence intervals or reducing the significance threshold α to account for the bias.

Secondly, covariate-adaptive randomization inference offers opportunities to develop stronger model-based guidance about the relative importance of the propensity score and the prognostic score in the construction of matched designs. Currently many design objectives proposed as bases for constructing matched sets, including mean balance, caliper matching, and multivariate distance matching have a heuristic element in the sense that it is not clear for which class of overall models for treatment and outcome they provide optimal results. While Kallus (2020) lays out a helpful overarching framework under sampling-based inference, unifying descriptions are not present for randomization-based inferences in matching. Covariate-adaptive inference provides important progress towards this goal by clarifying the impact of inexact propensity score matching on the operating characteristics of the ultimate estimate and associated test. Power analysis and design sensitivity calculations (Rosenbaum 2010b) based on the covariate-adaptive model, if developed, would provide valuable design-stage insight about how to properly use the propensity score in constructing the matches, and could provide more definitive guidance to users selecting among many Pareto optimal matches (Pimentel and Kelz 2020).

Thirdly, the evidence provided by our simulation study suggests that Z-dependence may play a limited but non-trivial role in failures of type I error control for matched randomization tests. While from a technical standpoint it may be sufficient toassume the matched pairs are generated independently by some model (as opposed to assuming a model on the original observations), this makes it difficult to develop formal understanding of how choices about matching influence inference. In particular, some statistical model on the original subjects of the study is needed to obtain the strong model-based guidance about the proper construction of matched sets discussed in the previous paragraph. Z-dependence differs fundamentally from propensity score discrepancies in the sense that it alters the support of the randomization distribution, not just the values of nonzero permutation probabilities, and new methods. The procedure outlined in Section 6.4, in which each permutation is checked to see if it produces the same match, works well for very small study sizes. Improving its computational performance seems possible; in particular, it should not be necessary to create an optimal match for each separate permutation, merely to check whether the current match is optimal or not, and this task may be easier to accomplish efficiently. We view developing such checks for commonly-used matching procedures as a promising future research direction.

Finally, while our development has focused exclusively on sharp null hypotheses under which both potential outcomes are known for all individuals, there is substantial practical interest in testing weak null hypotheses which allow for unknown heterogenous treatment effects and merely restrict the averages of such effects. Several threads of recent work have demonstrated that under mild conditions inference strategies developed to test sharp null hypotheses may also be valid tests for appropriately-chosen weak null hypotheses. An important task is to determine whether these ideas work for covariate-adaptive randomization inference. Notably, Caughey et al. (2021) showed that permutation tests of a sharp null may be reinterpreted as tests of a weak null describing maxima of treatment effects rather than averages. Since Caughey et al. place almost no restrictions on the permutation distribution and only mild conditions on the test statistic (all satisfied by the difference-in-means estimator we consider above), we anticipate that this approach should apply almost without modification to covariate-adaptive randomization tests. Fogarty (2020; 2022) describes sharp-null-style permutation tests and sensitivity analyses that are valid for the more traditional weak null hypothesis of zero average effect. These tests generally require studentized test statistics, and are much harder to construct for designs with matched sets containing more than two individuals. As such, more substantial extensions of our existing work are needed to see if the same ideas work under covariate-adaptive randomization inference.

Statement on conflicts of interest and data availability

The authors report no conflicts of interest. The welders data due originally to Costa et al. (1993) may be obtained via the R package `DOS2`, which is freely available at the Comprehensive R Archive Network (CRAN). The right heart catheterization data due originally to Connors et al. (1996) may be obtained via the R package `RBestMatch`. Although `RBestMatch` is not currently hosted by CRAN, archived versions of the package that contain the data are still freely available through the CRAN website at https://cran.r-project.org/src/contrib/Archive/RBestMatch/.

References

Abadie, A. and Imbens, G. W. (2006), “Large sample properties of matching estimators for average treatment effects,” *Econometrica*, 74, 235–267.

— (2011), “Bias-corrected matching estimators for average treatment effects,” *Journal of Business & Economic Statistics*, 29, 1–11.

Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018), "Doubly robust matching estimators for high dimensional confounding adjustment," *Biometrics*, 74, 1171–1179.

Austin, P. C. and Stuart, E. A. (2015), "Optimal full matching for survival outcomes: a method that merits more widespread use," *Statistics in medicine*, 34, 3949–3967.

Baiocchi, M. (2011), "Methodologies for observational studies of health care policy," Ph.D. thesis, University of Pennsylvania.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020), "The conditional permutation test for independence while controlling for confounders," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 175–197.

Branson, Z. and Bind, M.-A. (2019), "Randomization-based inference for Bernoulli trial experiments and implications for observational studies," *Statistical methods in medical research*, 28, 1378–1398.

Caughey, D., Dafoe, A., Li, X., and Miratrix, L. (2021), "Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects," *arXiv preprint arXiv:2101.09195*.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996), "The effectiveness of right heart catheterization in the initial care of critically III patients," *Jama*, 276, 889–897.

Costa, M., Zhitkovich, A., and Toniolo, P. (1993), "DNA-protein cross-links in welders: molecular implications," *Cancer research*, 53, 460–463.

Coudin, É. and Dufour, J.-M. (2020), "Finite-sample generalized confidence distributions and sign-based robust estimators in median regressions with heterogeneous dependent errors," *Econometric Reviews*, 39, 763–791.

Ding, P. and Dasgupta, T. (2016), "A potential tale of two-by-two tables from completely randomized experiments," *Journal of the American Statistical Association*, 111, 157–168.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

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

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

— (2022), "Testing weak nulls in matched observational studies," *Biometrics*.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2000), "Asymptotic separability in sensitivity analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 545–555.

Guo, K. and Rothenhäusler, D. (2023), "On the statistical role of inexact matching in observational studies," *Biometrika*, 110, 631–644.

Guo, W., Yin, M., Wang, Y., and Jordan, M. I. (2022), "Partial Identification with Noisy Covariates: A Robust Optimization Approach," *arXiv preprint arXiv:2202.10665*.

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

— (2009), "Propensity score matching to recover latent experiments: diagnostics and asymptotics," Tech. Rep. 486, University of Michigan.

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

Hodges Jr, J. L. and Lehmann, E. L. (1963), “Estimates of location based on rank tests,” *The Annals of Mathematical Statistics*, 34, 598–611.

Holland, P. W. (1986), “Statistics and causal inference,” *Journal of the American statistical Association*, 81, 945–960.

Jain, S., Rosenbaum, P. R., Reiter, J. G., Hill, A. S., Wolk, D. A., Hashemi, S., Fleisher, L. A., Eckenhoff, R., and Silber, J. H. (2022), “Risk of Parkinson’s disease after anaesthesia and surgery,” *British Journal of Anaesthesia*, 128, e268–e270.

Kallus, N. (2020), “Generalized optimal matching methods for causal inference.” *J. Mach. Learn. Res.*, 21, 62–1.

Kang, J. D. and Schafer, J. L. (2007), “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data,” *Statistical Science*, 22, 523–529.

Lehmann, E. L. (1959), *Testing statistical hypotheses*, John Wiley & Sons.

Lei, L. and Candès, E. J. (2021), “Conformal inference of counterfactuals and individual treatment effects,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83, 911–938.

Li, X. and Ding, P. (2017), “General forms of finite population central limit theorems with applications to causal inference,” *Journal of the American Statistical Association*, 112, 1759–1769.

Liu, H., Ren, J., and Yang, Y. (2022), “Randomization-based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2 K Factorial Experiments,” *Journal of the American Statistical Association*, 1–15.

Liu, H. and Yang, Y. (2020), “Regression-adjusted average treatment effect estimates in stratified randomized experiments,” *Biometrika*, 107, 935–948.

Luo, X., Dasgupta, T., Xie, M., and Liu, R. Y. (2021), “Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83, 777–797.

Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018), “Using standard tools from finite population sampling to improve causal inference for complex experiments,” *Journal of the American Statistical Association*, 113, 868–881.

Neyman, J. (1937), “Outline of a theory of statistical estimation based on the classical theory of probability,” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333–380.

Pashley, N. E., Basse, G. W., and Miratrix, L. W. (2021), “Conditional as-if analyses in randomized experiments,” *Journal of Causal Inference*, 9, 264–284.

Pimentel, S. D., Forrow, L. V., Gellar, J., and Li, J. (2020), “Optimal matching approaches in health policy evaluations under rolling enrolment,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183, 1411–1435.

Pimentel, S. D. and Kelz, R. R. (2020), “Optimal tradeoffs in matched designs comparing US-trained and internationally trained surgeons,” *Journal of the American Statistical Association*, 115, 1675–1688.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), “Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons,” *Journal of the American Statistical Association*, 110, 515–527.

Reinisch, J. M., Sanders, S. A., Mortensen, E. L., and Rubin, D. B. (1995), “In utero exposure to phenobarbital and intelligence deficits in adult men,” *Jama*, 274, 1518–1525.

Resa, M. d. l. A. and Zubizarreta, J. R. (2016), “Evaluation of Subset Matching Methods and Forms of Covariate Balance,” *Statistics in Medicine*, 35, 4961–4979.

Robins, J. M. and Wang, N. (2000), “Inference for imputation estimators,” *Biometrika*, 87, 113–124.

Rosenbaum, P. R. (1984), “Conditional permutation tests and the propensity score in observational studies,” *Journal of the American Statistical Association*, 79, 565–574.

— (1989), “Optimal matching for observational studies,” *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002a), “Covariance adjustment in randomized experiments and observational studies,” *Statistical Science*, 17, 286–327.

— (2002b), *Observational Studies*, New York, NY: Springer.

— (2010a), *Design of Observational Studies*, New York, NY: Springer.

— (2010b), “Design sensitivity and efficiency in observational studies,” *Journal of the American Statistical Association*, 105, 692–702.

— (2012), “Optimal Matching of an Optimally Chosen Subset in Observational Studies,” *Journal of Computational and Graphical Statistics*, 21, 57–71.

— (2018), “Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels,” *The Annals of Applied Statistics*, 12, 2312–2334.

Rosenbaum, P. R. and Krieger, A. M. (1990), “Sensitivity of two-sample permutation inferences in observational studies,” *Journal of the American Statistical Association*, 85, 493–498.

Rubin, D. B. (1980), “Comment on ‘Randomization analysis of experimental data: The Fisher randomization test’,” *Journal of the American Statistical Association*, 75, 591–593.

Sävje, F. (2021), “On the inconsistency of matching without replacement,” *Biometrika*.

Shaikh, A. M. and Toulis, P. (2021), “Randomization tests in observational studies with staggered adoption of treatment,” *Journal of the American Statistical Association*, 116, 1835–1848.

Shin, S. (2022), “Evaluating the effect of the matching grant program for refugees: An Observational Study Using Matching, Weighting, and the Mantel-Haenszel Test,” *Journal of Labor Research*, 43, 103–133.

Silber, J. H., Rosenbaum, P. R., Reiter, J. G., Hill, A. S., Jain, S., Wolk, D. A., Small, D. S., Hashemi, S., Niknam, B. A., Neuman, M. D., et al. (2020), “Alzheimer’s dementia after exposure to anesthesia and surgery in the elderly: a matched natural experiment using appendicitis,” *Annals of surgery*.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), “Multivariate matching and bias reduction in the surgical outcomes study,” *Medical care*, 1048–1064.

Stefanski, L. A. and Boos, D. D. (2002), “The calculus of M-estimation,” *The American Statistician*, 56, 29–38.

Stuart, E. A. (2010), “Matching methods for causal inference: A review and a look forward,” *Statistical Science*, 25, 1–21.

Tesema, G. A., Worku, M. G., Alamneh, T. S., Teshale, A. B., Yeshaw, Y., Alem, A. Z., Ayalew, H. G., Liyew, A. M., and Tessema, Z. T. (2023), “Estimating the impact of birth interval on under-five mortality in east african countries: a propensity score matching analysis,” *Archives of Public Health*, 81, 63.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019), “Conformal prediction under covariate shift,” *Advances in neural information processing systems*, 32.

Zhang, Y. and Zhao, Q. (2023), "What is a randomization test?" *Journal of the American Statistical Association*, 1–29.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019), "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

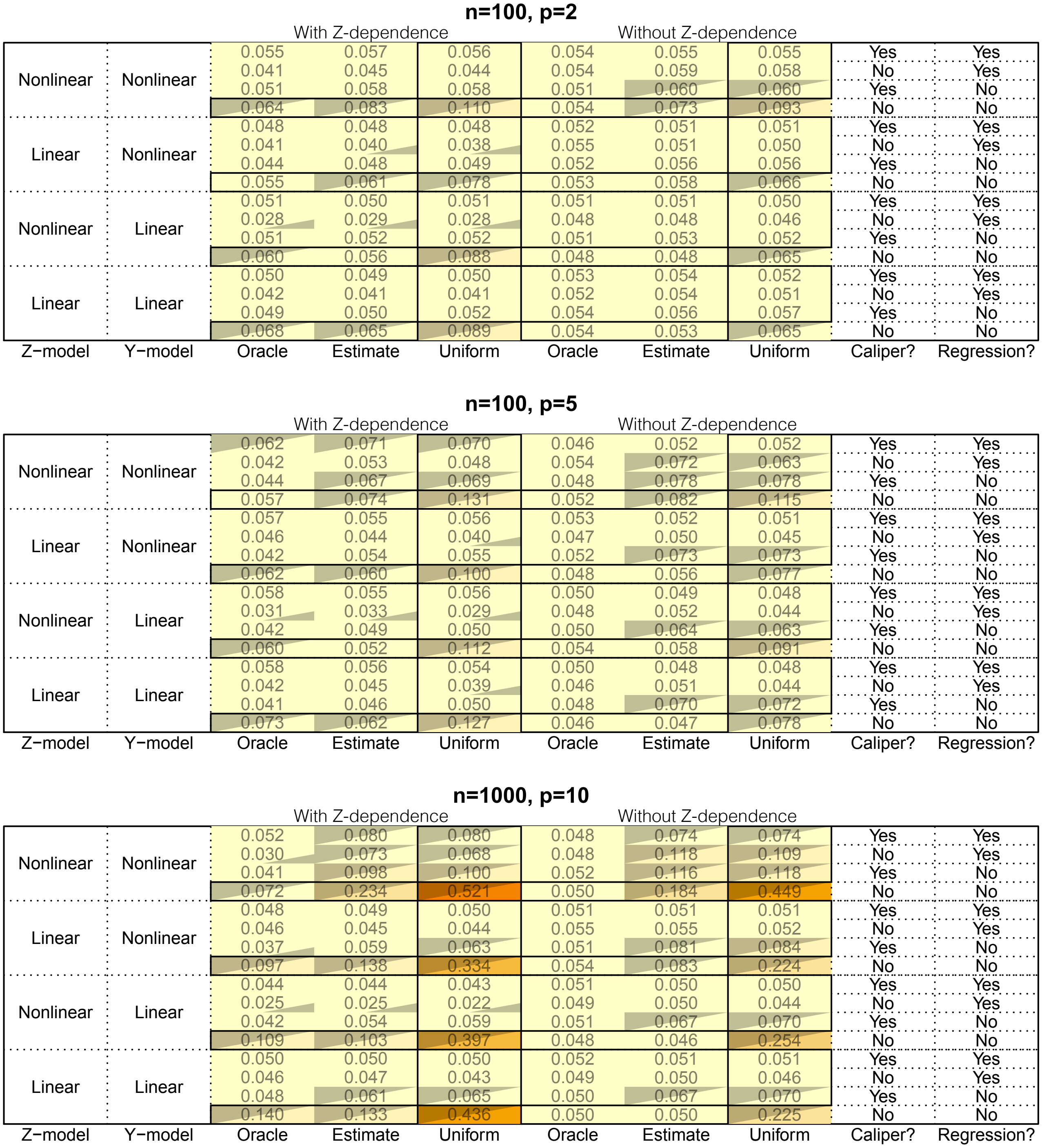
Zubizarreta, J. R. (2012), "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

Specification			Type I Error			Confidence interval length		
Z-model	Y-model	Regression?	Unmatched	Matched.	Z-dependence?	Unmatched	Matched.	Z-dependence?
				With	Without		With	Without
Linear	Linear	No	0.000	0.284	0.045	8.68	3.41	3.43
Linear	Linear	Yes	0.053	0.166	0.047	1.81	5.73	2.01
Linear	Nonlinear	No	0.001	0.037	0.051	7.70	2.11	3.22
Linear	Nonlinear	Yes	0.041	0.014	0.052	4.76	3.57	2.03
Nonlinear	Linear	No	0.000	0.258	0.056	8.85	3.39	5.22
Nonlinear	Linear	Yes	0.048	0.265	0.055	1.78	5.89	1.90
Nonlinear	Nonlinear	No	0.008	0.044	0.064	7.96	2.09	5.16
Nonlinear	Nonlinear	Yes	0.217	0.040	0.056	4.85	3.50	1.89

Table 1: Simulation results comparing matched and unmatched studies employing covariate-adaptive randomization inference. The first three columns describe distinct simulation settings (for $n = 100$ and $p = 2$ with other parameters given as described in Section 6.3). The middle three columns show Type I error computed as the proportion of rejections in a nominal level-0.05 test against a larger alternative across 5000 iterations; the first column gives the value for covariate-adaptive randomization inference in the full dataset without matching following Branson and Bind (2019), and the second two give values for covariate-adaptive randomization after matching (with and without Z-dependence) as detailed in Section 6.2. The final three columns show the average of two-sided confidence interval length (computed by inverting the two-sided version of the corresponding test) computed over the same set of iterations.

	Type I Error		
Regression?	Unmatched	Matched, with Z-dependence	Matched, no Z-dependence
No	0.135	0.072	0.064
Yes	0.139	0.059	0.058

Table 2: Simulation results comparing the robustness of matched and unmatched studies employing covariate-adaptive randomization inference to misspecification of the propensity score. Data are generated from a process with misspecification in both the propensity score and the outcome model as described in Section 6.4. Type I errors are computed as the proportion of rejections in a nominal level-0.05 test against a larger alternative across 5000 iterations.



n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.055	0.057	0.056	0.054	0.055	0.055	Yes	Yes
		0.041	0.045	0.044	0.054	0.059	0.058	No	Yes
		0.051	0.058	0.058	0.051	0.060	0.060	Yes	No
		0.064	0.083	0.110	0.054	0.073	0.093	No	No
Linear	Nonlinear	0.048	0.048	0.048	0.052	0.051	0.051	Yes	Yes
		0.041	0.040	0.038	0.055	0.051	0.050	No	Yes
		0.044	0.048	0.049	0.052	0.056	0.056	Yes	No
		0.055	0.061	0.078	0.053	0.058	0.066	No	No
Nonlinear	Linear	0.051	0.050	0.051	0.051	0.051	0.050	Yes	Yes
		0.028	0.029	0.028	0.048	0.048	0.046	No	Yes
		0.051	0.052	0.052	0.051	0.053	0.052	Yes	No
		0.060	0.056	0.088	0.048	0.048	0.065	No	No
Linear	Linear	0.050	0.049	0.050	0.053	0.054	0.052	Yes	Yes
		0.042	0.041	0.041	0.052	0.054	0.051	No	Yes
		0.049	0.050	0.052	0.054	0.056	0.057	Yes	No
		0.068	0.065	0.089	0.054	0.053	0.065	No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.062	0.071	0.070	0.046	0.052	0.052	Yes	Yes
		0.042	0.053	0.048	0.054	0.072	0.063	No	Yes
		0.044	0.067	0.069	0.048	0.078	0.078	Yes	No
		0.057	0.074	0.131	0.052	0.082	0.115	No	No
Linear	Nonlinear	0.057	0.055	0.056	0.053	0.052	0.051	Yes	Yes
		0.046	0.044	0.040	0.047	0.050	0.045	No	Yes
		0.042	0.054	0.055	0.052	0.073	0.073	Yes	No
		0.062	0.060	0.100	0.048	0.056	0.077	No	No
Nonlinear	Linear	0.058	0.055	0.056	0.050	0.049	0.048	Yes	Yes
		0.031	0.033	0.029	0.048	0.052	0.044	No	Yes
		0.042	0.049	0.050	0.050	0.064	0.063	Yes	No
		0.060	0.052	0.112	0.054	0.058	0.091	No	No
Linear	Linear	0.058	0.056	0.054	0.050	0.048	0.048	Yes	Yes
		0.042	0.045	0.039	0.046	0.051	0.044	No	Yes
		0.041	0.046	0.050	0.048	0.070	0.072	Yes	No
		0.073	0.062	0.127	0.046	0.047	0.078	No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.052	0.080	0.080	0.048	0.074	0.074	Yes	Yes
		0.030	0.073	0.068	0.048	0.118	0.109	No	Yes
		0.041	0.098	0.100	0.052	0.116	0.118	Yes	No
		0.072	0.234	0.521	0.050	0.184	0.449	No	No
Linear	Nonlinear	0.048	0.049	0.050	0.051	0.051	0.051	Yes	Yes
		0.046	0.045	0.044	0.055	0.055	0.052	No	Yes
		0.037	0.059	0.063	0.051	0.081	0.084	Yes	No
		0.097	0.138	0.334	0.054	0.083	0.224	No	No
Nonlinear	Linear	0.044	0.044	0.043	0.051	0.050	0.050	Yes	Yes
		0.025	0.025	0.022	0.049	0.050	0.044	No	Yes
		0.042	0.054	0.059	0.051	0.067	0.070	Yes	No
		0.109	0.103	0.397	0.048	0.046	0.254	No	No
Linear	Linear	0.050	0.050	0.050	0.052	0.051	0.051	Yes	Yes
		0.046	0.047	0.043	0.049	0.050	0.046	No	Yes
		0.048	0.061	0.065	0.050	0.067	0.070	Yes	No
		0.140	0.133	0.436	0.050	0.050	0.225	No	No

Figure 1: Type I error results for uniform and covariate-adaptive inference across multiple simulation settings. Each of the three tables corresponds to a separate dataset size. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models. Numbers give type I error rates with colors associated to their magnitude; triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05 (under a Bonferroni correction scaled to the number of results across the entire figure), with a large upper triangle indicating a positive z-statistic and a small lower triangle indicating a negative z-statistic.

n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	2.013	2.024	2.025	2.014	2.025	2.026	Yes	Yes
		1.915	1.938	1.949	1.917	1.940	1.951	No	Yes
		2.099	2.123	2.124	2.101			Yes	No
					1.998			No	No
Linear	Nonlinear	2.007	2.016	2.017	2.007	2.015	2.016	Yes	Yes
		1.901	1.919	1.928	1.898	1.916	1.926	No	Yes
		2.096	2.109	2.110	2.098	2.111	2.112	Yes	No
		1.971			1.969	1.999		No	No
Nonlinear	Linear	2.036	2.040	2.041	2.034	2.038	2.039	Yes	Yes
		1.931	1.923	1.943	1.930	1.922	1.943	No	Yes
		2.130	2.135	2.137	2.126	2.132	2.133	Yes	No
			2.045		2.057	2.045		No	No
Linear	Linear	2.027	2.031	2.032	2.036	2.039	2.040	Yes	Yes
		1.906	1.899	1.917	1.911	1.903	1.921	No	Yes
		2.118	2.122	2.124	2.123	2.128	2.130	Yes	No
					1.994	1.985		No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear				1.903	1.940	1.941	Yes	Yes
		1.806	1.816	1.868	1.802			No	Yes
		2.139			2.137			Yes	No
		1.999			1.995			No	No
Linear	Nonlinear	1.877	1.904	1.906	1.877	1.904	1.906	Yes	Yes
		1.784	1.782	1.830	1.773	1.772	1.819	No	Yes
		2.147	2.190	2.192	2.141			Yes	No
					1.983	1.991		No	No
Nonlinear	Linear	1.923	1.947	1.949	1.917	1.942	1.944	Yes	Yes
		1.813	1.772	1.853	1.815	1.775	1.855	No	Yes
		2.177	2.212	2.216	2.166			Yes	No
			2.001		2.056	2.002		No	No
Linear	Linear	1.901	1.923	1.925	1.896	1.918	1.920	Yes	Yes
		1.771	1.732	1.804	1.773	1.735	1.806	No	Yes
		2.162	2.194	2.197	2.157			Yes	No
					1.995	1.947		No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.629			0.629			Yes	Yes
		0.614			0.613			No	Yes
		0.643			0.643			Yes	No
					0.632			No	No
Linear	Nonlinear	0.624	0.630	0.630	0.624	0.630	0.630	Yes	Yes
		0.604	0.613	0.622	0.603	0.613	0.621	No	Yes
		0.641	0.647		0.641			Yes	No
					0.628			No	No
Nonlinear	Linear	0.627	0.630	0.630	0.627	0.629	0.630	Yes	Yes
		0.619	0.616	0.636	0.619	0.617	0.636	No	Yes
		0.641	0.644	0.644	0.641			Yes	No
					0.650	0.647		No	No
Linear	Linear	0.624	0.626	0.626	0.624	0.626	0.626	Yes	Yes
		0.603	0.601	0.616	0.603	0.601	0.616	No	Yes
		0.638			0.638			Yes	No
					0.631	0.628		No	No

Figure 2: Average confidence interval length for uniform and covariate-adaptive inference across multiple simulation settings for cases with approximate control of type I error at 0.05 or less. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models.

	African-American	Age	Smoker	PS
Pair 1	0.00	-6.00	0.00	0.12
Pair 2	0.00	-3.00	0.00	0.05
Pair 3	0.00	4.00	0.00	0.05
Pair 4	0.00	-8.00	0.00	0.16
Pair 5	0.00	0.00	0.00	0.00
Pair 6	0.00	3.00	0.00	-0.06
Pair 7	0.00	-3.00	0.00	0.06
Pair 8	0.00	-2.00	0.00	0.17
Pair 9	0.00	3.00	0.00	-0.06
Pair 10	0.00	-5.00	0.00	0.06
Pair 11	0.00	-3.00	0.00	0.06
Pair 12	0.00	-5.00	0.00	0.10
Pair 13	0.00	-5.00	0.00	0.11
Pair 14	0.00	-5.00	0.00	0.10
Pair 15	0.00	0.00	0.00	0.00
Pair 16	0.00	-4.00	0.00	0.08
Pair 17	0.00	1.00	0.00	-0.02
Pair 18	0.00	1.00	0.00	-0.01
Pair 19	-1.00	7.00	0.00	0.04
Pair 20	0.00	-5.00	0.00	0.10
Pair 21	0.00	4.00	0.00	0.05
Average	-0.05	-1.48	0.00	0.06

Table 3: Treated- control differences in matched pairs in the welders dataset.

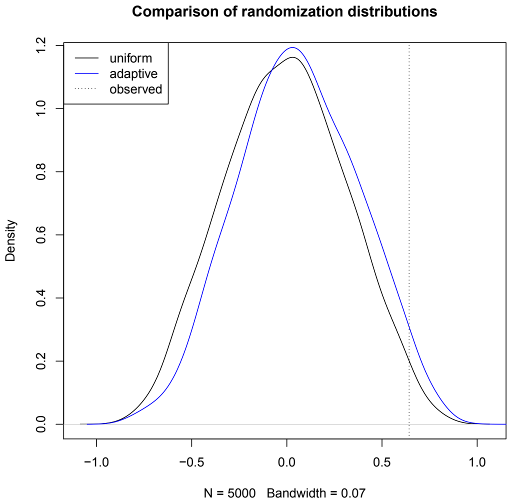


Figure 3: Smoothed densities for the uniform randomization distribution of the difference in means statistic and the covariate-adaptive randomization distribution in the welders dataset, with the value of the observed statistic.

	Before Matching	Caliper Match	Subset Match
Est. Propensity Score	1.254	0.002	0.033
Acute Physiology Score	0.501	0.016	0.068
Mean Blood Pressure	-0.455	-0.026	-0.058
PaO2/(.01*FiO2)	-0.433	0.001	0.030
Serum Creatinine	0.270	0.021	0.044
Hematocrit	-0.269	0.001	-0.008
Weight (kg)	0.256	0.002	-0.017
PaCO2	-0.249	0.004	-0.013
MOSF w/Sepsis (secondary)	0.230	-0.036	-0.014
Albumin	-0.230	-0.018	-0.021
Predicted Survival	-0.198	-0.031	-0.049
MOSF w/Sepsis (primary)	0.172	0.000	0.000
Respiration Rate	-0.165	-0.002	-0.011
Heart Rate	0.147	0.027	0.000
Bilirubin	0.145	-0.013	0.010
Heart disease	0.139	0.011	0.014
MOSF w/Malignancy (secondary)	-0.135	0.009	-0.017
Respiratory Disease	-0.128	-0.007	-0.020
Serum pH	-0.120	-0.002	-0.044
Missing ADL Score	0.117	-0.001	0.005
SUPPORT Coma Score	-0.110	0.045	0.022
Neurological disease	-0.108	0.006	0.001
Serum Sodium	-0.092	-0.011	-0.001
Years of Education	0.091	0.022	0.024
Sepsis	0.091	0.012	0.017

Table 4: Standardized differences in means before matching and under two different matches for the 25 variables with largest initial imbalance in the right-heart catheterization dataset.

	Caliper Match	Subset Match
Avg. Propensity Score Discrepancy	0.016	0.166
Max Propensity Score Discrepancy	0.030	0.672
Prop. Matched Exactly, Male Sex	0.351	0.297
Avg. Discrepancy, Age	13.469	10.892
Avg. Discrepancy, Mean Blood Pressure	27.809	25.506
Avg. Discrepancy, Heart Rate	32.677	28.858
Avg. Discrepancy, Respiration Rate	11.614	10.426
Avg. Discrepancy, PaO2/(.01*FiO2)	76.651	75.882

Table 5: Matched pair quality under two different matches for selected variables the right-heart catheterization dataset.

	Caliper match		Subset match	
	Uniform	Covariate-Adaptive	Uniform	Covariate-Adaptive
Risk difference	0.074	0.073	0.068	0.062
Lower conf. limit	0.043	0.041	0.038	0.034
Upper conf. limit	0.106	0.104	0.098	0.089
Threshold Γ	1.12	1.12	1.11	1.12
Risk difference, OLS-adjusted	0.051	0.051	0.055	0.054
Lower conf. limit, OLS-adjusted	0.022	0.022	0.027	0.028
Upper conf. limit, OLS-adjusted	0.081	0.081	0.083	0.079
Threshold Γ, OLS-adjusted	1.06	1.06	1.08	1.09

Table 6: Outcome analysis for 30-day mortality in the right-heart catheterization data. Results are shown for both the caliper match and the subset match, with and without regression adjustment, and using both uniform and covariate-adaptive inference.

List of figure captions

1. Type I error results for uniform and covariate-adaptive inference across multiple simulation settings. Each of the three tables corresponds to a separate dataset size. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models. Numbers give type I error rates with colors associated to their magnitude; triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05 (under a Bonferroni correction scaled to the number of results across the entire figure), with a large upper triangle indicating a positive z-statistic and a small lower triangle indicating a negative z-statistic.

2. Average confidence interval length for uniform and covariate-adaptive inference across multiple simulation settings for cases with approximate control of type I error at 0.05 or less. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models.

3. Smoothed densities for the uniform randomization distribution of the difference in means statistic and the covariate-adaptive randomization distribution in the welders dataset, with the value of the observed statistic.